

**ASSESSMENT OF MATHEMATICAL PROBLEM SOLVING: ADDITION AND SUBTRACTION IN
ARITHMETIC WORD PROBLEM SOLVING**

Farhan Azim

orcid.org/0000-0001-8011-1090

Submitted in total fulfilment of the requirements of the degree of
Doctor of Philosophy

Melbourne Graduate School of Education
The University of Melbourne

2020

Abstract

Researchers consider mathematical problem solving (MPS) to be a vehicle for teaching and reinforcing mathematical knowledge, solving real life problems, and helping to meet everyday challenges. Like many other countries in the world, Bangladesh has given great importance to MPS in its primary curriculum. The country has achieved well in quantitative parameters of primary education (e.g., net enrolment rate, completion rate, etc.) but there are concerns about the quality of learning. This is particularly the case for primary mathematics, as exhibited in different examinations. One of the issues with mathematics teaching and learning is the lack of formative assessment materials that teachers can utilise to diagnose students' learning deficiencies and use as a basis for designing appropriate remedial teaching. Furthermore, there is a paucity of research on mathematics assessment at the primary level in Bangladesh that has addressed the issues related to quality of learning. It is the aim of this research to develop and validate assessment materials using the latest developments in psychometrics that can (a) be used for this purpose, and (b) serve as a model for further relevant work.

Arithmetic word problem solving (AWPS) is an area within MPS that is often used to help children learn how to apply the formal mathematical knowledge and skills they acquire at school in real-life situations. Although AWPS plays an important role in understanding mathematics, it is an area where students experience great difficulty. It is a complex activity that requires a series of cognitive processes, but current assessment of AWPS in Bangladesh relies exclusively on giving a single score based on the solution being correct or incorrect. There is limited opportunity for understanding students' difficulties in the component cognitive processes. In the global context, empirical studies exploring the cognitive processes involved in solving arithmetic word problems are scarce. In the Bangladeshi education context, no studies were found that focused on AWPS.

This research endeavoured to develop assessment materials that will enable teachers to identify specific difficulties students face while solving AWPS. The research objectives were to (a) develop a theoretical framework for arithmetic word problem solving using an iterative process (as discussed by Wilson, 2005) by (b) constructing tasks aligned with the framework, (c) trying them out in both small- and large-scale studies, (d) investigating the results using qualitative and quantitative methods, and (e) evaluating the success of the process by investigating the validity and reliability of the results. Only addition and subtraction operations for AWPS were considered to limit the scope of the study. Based on review of existing literature, a framework for assessing AWPS was developed. This framework theorises five cognitive elements/strands of AWPS: comprehension of the problem, mapping conceptual relations to a numerical representation, selecting an appropriate strategy for solution, performing the calculations, and verifying the solution and communicating the results. The validity and reliability of the framework were investigated by developing and administering a test (with items representing the different processes). The test was administered in two phases in Bangladesh: a pilot study with 62 students (from one school) and a final study with 808 students (from nine schools). Grade 4 and grade 5 students were selected through convenience sampling for both iterations of data collection. Four

different types of validity evidence (based on test content, response processes, internal structure, and relations with other variables) were collected and analysed.

In addition to using test design and specification, an expert opinionnaire was used to collect validity evidence based on test content. Analysis of this data showed that overall, the experts' association of strands with items agreed with what was hypothesised, that is, they nominated items for strands they were hypothesised to be representing in this research. Cognitive interviews were conducted with 20 respondents from three schools to collect validity evidence based on the response processes. Data from these cognitive interviews were analysed using a qualitative method called 'reduction and synthesis' (Miller et al., 2014) and showed that the items were mostly working as intended. An Item Response Theory (IRT) approach was adopted to evaluate the internal structure of the test. A dichotomous Rasch analysis was used. The analysis showed that the item difficulties were appropriately distributed, and that the item fit statistics suggested that almost all the items were within a reasonable range (weighted MNSQ 0.88 to 1.14) and did not show any extreme over-fit or under-fit. The fit t statistics for the items also indicated a good fit. Overall, the data showed good adherence to the Rasch model and can be considered to be demonstrating reasonably good evidence of validity in terms of internal structure of the test. Two types of data were collected with the aim of establishing validity arguments based on the AWPS test's relation with other variables: (a) scores from six textbook items included in the test that was administered and (b) students' test scores for Bangla and Mathematics from their schools. These scores were positively correlated with students' performance on the test conducted in this study, and hence showed good evidence of validity in relation to other variables. Test reliability was calculated using both IRT and CTT measures. The WLE person separation reliability was 0.81 and the EAP/PV reliability was 0.82. The Cronbach's Alpha value was 0.82.

The findings of this study can inform the inclusion of appropriate assessment techniques in the primary mathematics curriculum in Bangladesh, and the development of professional development materials for teachers using the assessments. Teachers of mathematics could utilise the framework to design their own tests that would help them to diagnose students' learning deficiencies and design appropriate remedial teaching. In addition, future research can replicate this study in different settings to refine the AWPS framework and use it in different contexts. Furthermore, the methodology used in this study could be used in future research to develop assessment materials for other areas of mathematics including other areas of problem solving (e.g., arithmetic word problem solving where other operations like multiplication or division are used).

Declaration

This is to certify that:

- 1) This thesis comprises only my original work towards the degree of Doctor of Philosophy;
- 2) Due acknowledgement has been made in the text to all other material used; and
- 3) The thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies, and appendices.

Farhan Azim

Preface

This research was financially supported by Melbourne International Fee Remission Scholarship (MIFRS) and Melbourne International Research Scholarship (MIRS).

The final draft of this thesis was proofread by Dr Colleen Keane, who provided grammatical and semantic feedback, but preserved my meaning, style, organisation and voice.

Acknowledgements

I would like to express my deepest sense of gratitude to my principal supervisor Professor Mark Wilson for his exceptional inspiration, wisdom, and encouragement throughout my PhD journey. I would also like to express my heartfelt appreciation to his co-supervisor Dr Esther Chan for her scholarship, guidance and constructive critiques of this research work. I am indebted to Professor Janet Clinton, the chair of my advisory committee, for her valuable insights about the study. This research also owes much to Professor David Clarke, who was a co-supervisor before passing away earlier this year. His contributions were instrumental in developing the framework and the items used in this study. He is dearly missed.

I would like to acknowledge the technical expertise of Dr Zhonghua Zhang, Dr Masa Pavlovic, and Dr Cuc Nguyen, who helped me to learn about psychometrics throughout this research. It was a privilege to have the kind support of Dr Bruce Beswick, Dr Rebekah Luo, Ms Hilary Slater, Ms Bernadette Coles-Janess and other colleagues of the Assessment Research Centre (ARC) and Dr Mahtab Janfada, Dr Daniela Acquaro, Ms Rosie Yasmin, Dr Vicki Steinle, Dr Lynda Ball and other colleagues of the MGSE at various stages of this PhD.

This work would not have been possible without the collegiality and friendship of my PhD colleagues at the ARC. I want to thank Dr Emily White, Dr Jane Strickland, Dr Toshiko Kamei, Dr Nives Nibali, Dr Dennis Danipong, Ms Pik Yen Lim, Dr Lingfen Zhang, Dr Shafqat Rahman, Ms Jolanta Szymakowski, Ms Lan Anh Nguyen Khoa, Ms Thi Nhan, and Ms Zhihong Sun for the lively discussions and the opportunities of growth we shared.

I also want to thank colleagues from Bangladesh for their support in this research. Professor Hosne Ara Begum acted as the local mentor for the fieldwork despite her busy schedule. Associate Professor Shah Shamim Ahmed, Mr Rakib Rokon, Ms Iffat Naomee and Ms Roti Chakraborty helped in locating research literature. Mr Saiful Islam, Ms Israt Arefin and Ms Sharmin Hena helped with the translations and the data collection. Dr Fatema Taj and Ms Deya Chakraborty provided valuable feedback at the initial stages of the instrument development.

I want to convey my sincerest appreciation to the primary school teachers, parents, students and experts from Bangladesh who helped in the data collection for this research.

Finally, I want to thank my family for all the sacrifices they have made and the support they have provided throughout my education journey.

Table of Contents

Abstract	ii
Declaration	iv
Acknowledgements	vi
Table of Contents	vii
List of Tables	xii
List of Figures	xiv
List of Acronyms and Abbreviations	xv
Chapter 1: Background of the study	1
1.1 The Educational Context of Bangladesh	2
1.2 Mathematics in Bangladesh’s primary curriculum	6
1.3 Mathematics proficiency of primary students in Bangladesh	11
1.3.1 Government administered large-scale assessment.....	12
1.3.2 NGO administered large scale assessment.....	17
1.3.3 Small-scale research studies	18
1.4 Primary teachers in Bangladesh.....	21
1.5 Rationale of this study	22
1.6 Research objectives and research questions	26
1.7 Structure of the thesis	27
Chapter 2: Review of literature	31
2.1 What is a Problem?	31
2.2 Problem Solving	32
2.3 Historical background of concepts of problem solving.....	34
2.4 Domain specific problem solving	37
2.5 Mathematical problem solving	39
2.6 Teaching of mathematical problem solving.....	43
2.7 Specific area within Mathematical Problem Solving: Arithmetic Word Problem Solving	45
2.8 Developing a theoretical framework for arithmetic word problem solving	49
2.8.1 Theories of Mathematical Problem Solving.....	50
2.8.2 Theories of Arithmetic Problem Solving	56
2.8.3 Strategies in solving arithmetic word problems	57

2.8.4 Error analysis in written problem solving	59
2.9 A theoretical framework for assessing Arithmetic Word Problem Solving	60
2.10 Approaches to validity and reliability	65
2.10.1 Validity	65
2.10.2 Developing a coherent validity argument based on different types of evidence ..	75
2.10.3 Reliability.....	75
Chapter 3: Methodology and methods.....	81
3.1 Research philosophy	82
3.1.1 Ontology.....	85
3.1.2 Epistemology.....	87
3.1.3 Axiology.....	88
3.1.4 Pragmatism as the research paradigm of this study	89
3.2 Research Methodology	91
3.2.1 Mixed methods as a research methodology.....	93
3.3 Data collection instruments.....	95
3.3.1 Student test.....	97
3.3.2 Students' cognitive interviews.....	111
3.3.3 Expert opinionnaire.....	116
3.3.4 Students' Mathematics and Bangla test scores from the school	116
3.3.5 School background questionnaire	117
3.3.6 Translation of instruments.....	117
3.4 Participants of the study	120
3.4.1 Population.....	125
3.4.2 Selection of sample for the pilot study.....	126
3.4.3 Selection of sample for the final data collection	128
3.5 Administering the data collection instruments	134
3.5.1 Data collection for pilot	134
3.5.2 Final data collection	137
3.6 Data management	144
3.6.1 Data management for pilot.....	144
3.6.2 Data management for final data collection	145
3.6.3 Date entry and data cleaning.....	149
3.7 Data analysis:	151

3.7.1 Psychometric analysis of the test in piloting and final data collection.....	152
3.7.2 Analysis of other quantitative data.....	158
3.7.3 Analysis of qualitative data from students' cognitive interviews.....	159
3.8 Ethical considerations of this research	164
Chapter 4: The pilot study.....	165
4.1 Data analysis for the pilot test.....	167
4.1.1 Item analysis	169
4.1.2 Item fit.....	177
4.1.3 Response fit.....	180
4.1.4 Item-total correlation	183
4.1.5 Analysis of test reliability	187
4.1.6 Person standard errors	187
4.1.7 Students' post-test cognitive interviews	188
4.2 Analysis of validity evidence based on relations with other variables	190
4.3 Item revision based on pilot data	192
Chapter 5: Analysis of quantitative data	196
5.1 Analysis of validity evidence based on internal structure: Students' AWPS test	196
5.1.1 Frequency analysis of the AWPS test data.....	198
5.1.2 Item analysis	201
5.1.3 Item fit.....	209
5.1.4 Response fit.....	213
5.1.5 Item-total correlation	219
5.2 Model fit: Unidimensional vs Multidimensional.....	222
5.3 Analysis of test reliability	224
5.4 Analysis of validity evidence based on relations with other variables	226
5.4.1 Comparison of AWPS test and textbook item score.....	228
5.4.2 Comparison of AWPS test and school test scores	230
5.5 Analysis of validity evidence based on test content: Expert opinionnaire	234
Chapter 6: Analysis of qualitative data.....	250
6.1 Strand 1 - Comprehension of the problem.....	252
6.1.1 Summary of analysis for Strand 1 items	259
6.2 Strand 2 - Mapping conceptual relations to a numerical representation	259

6.2.1 Summary of analysis for Strand 2 items	269
6.3 Strand 3 - Selecting an appropriate strategy for solution.....	270
6.3.1 Summary of analysis for Strand 3 items	275
6.4 Strand 4 - Performing the calculations on the numbers.....	276
6.4.1 Summary of analysis for Strand 4 items	282
6.5 Strand 5 - Verifying the solution and communicating the results	283
6.5.1 Summary of analysis for Strand 5 items	293
Chapter 7: Findings and discussions	295
7.1 Reflections on the study	295
7.1.1 What were the objectives of this study and were they fulfilled?.....	295
7.1.2 Significance of the study	307
7.1.3 Successes and limitations of the study	308
7.1.4 How could this work be extended, and what could be done differently in future studies?	310
7.1.5 What changes could be brought to the methodology of this study?	312
7.1.6 What other types of analysis could be conducted?.....	313
7.2 Implications of this research for Bangladesh.....	314
7.2.1 Policymakers	314
7.2.2 Mathematics teachers	317
Bibliography.....	320
Appendices	334
Appendix 1.1: Assessment of word problem solving in the Teacher Editions (grades three to five)	335
Appendix 3.1: Construct map and item content specifications for the pilot study items....	339
Appendix 3.2: Construct map and item content specifications for the final test items.....	344
Appendix 3.3: Pilot study test booklet (English)	349
Appendix 3.4: Pilot study test booklet (Bangla)	359
Appendix 3.5: Answer key for the pilot study test	371
Appendix 3.6: Final data collection test booklet (English)	372
Appendix 3.7: Final data collection test booklet (Bangla)	382
Appendix 3.8: Answer key for the test administered in the final data collection	394
Appendix 3.9: Expert opinionnaire (English).....	395
Appendix 3.10: Expert opinionnaire (Bangla)	410

Appendix 3.11: Plain language statement (PLS) and consent form for head teachers	430
Appendix 3.12: School background questionnaire (English)	433
Appendix 3.13: School background questionnaire (Bangla)	435
Appendix 3.14: Revision of items during translation	437
Appendix 3.15: Plain language statement (PLS) and consent form for parents (English) ...	440
Appendix 3.16: Plain language statement (PLS) and consent form for parents (Bangla)....	443
Appendix 3.17: Plain language statement (PLS) for students (English)	446
Appendix 3.18: Plain language statement (PLS) for students (Bangla)	447
Appendix 3.19: Assent form for students (English)	449
Appendix 3.20: Assent form for students (Bangla)	450
Appendix 3.21: Students' cognitive interview fieldnote template for the pilot study (English)	451
Appendix 3.22: Students' cognitive interview fieldnote template for the pilot study (Bangla)	470
Appendix 3.23: Students' cognitive interview fieldnote template for the final study (English)	489
Appendix 3.24: Students' cognitive interview fieldnote template for the final data collection (Bangla)	505
Appendix 3.25: Plain language statement (PLS) and consent form for experts	521
Appendix 3.26: Codebook for the data collection in the final phase	524
Appendix 4.1: WLE for the pilot study respondents (sorted by fit)	542
Appendix 6.1: Ability groups of cognitive interview respondents	544

List of Tables

Table	Page
<i>Table 2.1 Connection between the theories discussed and the hypothesised AWPS framework</i>	60
<i>Table 3.1 Sources of Validity Evidence</i>	96
<i>Table 3.2 Sources of Textbook Items Included in the Test Booklet</i>	109
<i>Table 3.3 Number of Respondents from Each School for AWPS Test in the Final Data Collection</i>	130
<i>Table 3.4 Respondent Sample Distribution for Expert Opinionnaire</i>	131
<i>Table 3.5 Distribution of Students Selected for the Cognitive Interviews across Different Schools in the Final Phase Data Collection</i>	133
<i>Table 3.6 Codes for Open Ended Items (03, 09, 15, 21, 27, 33)</i>	147
<i>Table 3.7 Data Entry Error Check Results</i>	150
<i>Table 4.1 Item fit t Statistics for the Pilot Study</i>	179
<i>Table 4.2 Person Fit Statistics for the Pilot Study</i>	180
<i>Table 4.3 Response Pattern for Respondent 29 (SRID 14030) when Items are Sorted by Empirically Attained Item Difficulty</i>	181
<i>Table 4.4 Response Pattern for Respondent 60 (SRID 15028) when Items are Sorted by Empirically Attained Item Difficulty</i>	182
<i>Table 4.5 Response Pattern for Respondent 47 (SRID 14015) and student 56 (SRID 14024) when Items are Sorted by Empirically Attained Item Difficulty</i>	183
<i>Table 4.6 Item-Rest Correlation of Items in the Pilot Study Test</i>	183
<i>Table 4.7 Item Difficulty of Strategy Selection Items in the Pilot Study Test</i>	192
<i>Table 4.8 Revising Item Specifications for the Strategy Selection Items</i>	192
<i>Table 4.9 Item Difficulty of the Performing Calculation Items in the Pilot Study Test</i>	193
<i>Table 4.10 Revising Item Specifications for the Performing Calculation Items</i>	194
<i>Table 5.1 Distribution of Items in the Final Test Booklet</i>	197
<i>Table 5.2 Item Fit T Statistics</i>	212
<i>Table 5.3 Person Fit Statistics for Respondents Showing Underfit (Fit Statistics 2.00 and Above)</i>	214

Table 5.4 <i>Response Pattern for Respondent 277 (SRID 544114) when Item Sequence Is Revised Based on Empirically Obtained Item Difficulty</i>	215
Table 5.5 <i>Response Pattern for Respondent 179 (SRID 35057) when Item Sequence Is Revised Based on Empirically Obtained Item Difficulty</i>	216
Table 5.6 <i>Response Pattern for Respondent 329 (SRID 55036) when Item Sequence Is Revised Based on Empirically Obtained Item Difficulty</i>	216
Table 5.7 <i>Comparison of Fit between Unidimensional and Multidimensional Model</i>	223
Table 5.8 <i>Estimated Correlations among the Dimensions in the Multidimensional Model</i>	224
Table 5.9 <i>Reliability Coefficients for the Different Dimensions of the Multidimensional Model</i>	225
Table 5.10 <i>Correlation between Students' Scaled Score in AWPS Test and Textbook Item Scores</i>	228
Table 5.11 <i>Correlations between Students' Textbook Item Scores and WLE estimates for the Dimensions of the Multidimensional Model</i>	229
Table 5.12 <i>Correlation Between Students' Scaled Score in the AWPS Test and their Mathematics Test Score in School Examination</i>	230
Table 5.13 <i>Correlations between Students' Mathematics Test Score in School Examination and their WLE estimates for the Dimensions of the Multidimensional Mode</i>	231
Table 5.14 <i>Correlation Between Students' Scaled Score in the AWPS Test and their Bangla Test Score in School Examination</i>	232
Table 5.15 <i>Correlations between Students' Bangla Test Score in School Examination and their WLE estimates for the Dimensions of the Multidimensional Model</i>	233
Table 7.1 <i>Validity Evidence Collected for this Research</i>	297

List of Figures

Figure	Page
<i>Figure 4.1.</i> Missing response percentage per item for the pilot study	168
<i>Figure 4.2.</i> The Wright map for the pilot study test (all strands together)	170
<i>Figure 4.3.</i> The Wright map for the pilot study test (strands separated)	173
<i>Figure 4.4.</i> Weighted MNSQ of pilot study items	178
<i>Figure 4.5.</i> Standard error of measurement for the pilot AWPS test	188
<i>Figure 5.1.</i> Percentages of missing response for each item (textbook items excluded)	199
<i>Figure 5.2.</i> The Wright map for final data collection test (strands separated)	201
<i>Figure 5.3.</i> Weighted MNSQ (infit) of items	211
<i>Figure 5.4.</i> Kidmap for respondent 329 (SRID-55036)	218
<i>Figure 5.5.</i> Item-rest correlations of items in the final test	219
<i>Figure 5.6.</i> Item-rest correlations of items in the final test when item 35 is deleted.....	221
<i>Figure 5.7.</i> Standard Error of Measurement for the Final AWPS Test	226
<i>Figure 5.8.</i> Expert opinion: Strand 1 – Comprehension of the Problem items	236
<i>Figure 5.9.</i> Expert opinion: Strand 2 – Mapping Conceptual Relations to a Numerical Representation items	238
<i>Figure 5.10.</i> Expert opinion: Strand 3 – Selecting an Appropriate Strategy items	240
<i>Figure 5.11.</i> Expert opinion: Strand 4 – Performing Calculations items	242
<i>Figure 5.12.</i> Expert opinion: Strand 5 – Verifying the Solution and Communicating the Results items (set 1)	244
<i>Figure 5.13.</i> Expert opinion: Strand 5 – Verifying the Solution and Communicating the Results items (set 2)	246

List of Acronyms and Abbreviations

ADB	Asian Development Bank
ACER	Australian Council for Educational Research
AIR	American Institutes for Research
APSC	Annual Primary School Census
ASPR	Annual Sector Performance Report
BANBEIS	Bangladesh Bureau of Educational Information and Statistics
CAMPE	Campaign for Popular Education
DPE	Directorate of Primary Education
EECE	Ebtedayee Education Completion Examination
EFA	Education for All
EW	Education Watch
HSC	Higher Secondary Certificate
IED	Institute of Educational Development
MDG	Millennium Development Goals
MoPME	Ministry of Primary and Mass Education
MWCA	Ministry of Women and Children Affairs
NAEM	National Academy for Educational Management
NAPE	National Academy of Primary Education
NCTB	National Curriculum and Textbook Board
NGO	Non-Government Organization

NSA	National Student Assessment
OECD	Organization for Economic Co-operation and Development
PECE	Primary Education Completion Examination
PEDP	Primary Education Development Program
PISA	Programme for International Student Assessment
PTI	Primary Teachers Training Institute
SDG	Sustainable Development Goals
SSC	Secondary School Certificate
SPSS	Statistical Package for the Social Sciences

Chapter 1: Background of the study

Problem solving has taken up an important role in many education systems. A number of countries around the world consider achieving problem solving competency to be a key objective of their educational programs. Acquiring increased levels of problem solving skill gives learners a better platform for learning in later stages of life, makes them successful participants in the society, and enhances their ability of conducting personal activities (OECD, 2012). As educators are interested to improve students' ability to solve problems, problem solving has become fundamental in education (Mayer & Wittrock, 2009) and teachers often consider developing students' problem solving ability as a primary objective of education (Lester & Kehle, 2003). The education system in Bangladesh also emphasises developing students' capacity of problem solving. Even though it is not always well articulated within the policy documents, researchers have highlighted how it is important for the students in Bangladesh to achieve this skill (e.g., Hossain, 2015).

This study focused on developing a demonstration assessment system for one specific area of problem solving – Arithmetic Word Problem Solving (AWPS). The aim was to develop exemplary assessment materials that will provide a model for what is needed for sound formative assessment and hence lay the foundations for future professional development that will enable teachers to identify specific difficulties students face while solving arithmetic word problems. To give an overview of this research, the next sections will describe the education context of Bangladesh, rationale for this study, and this study's aims and objectives.

1.1 The Educational Context of Bangladesh

The education system in Bangladesh is comprised of five years of compulsory primary education (grade 1 to grade 5), seven years of secondary education, and higher (tertiary) education. The seven-year cycle of secondary school is further divided into junior-secondary (grades 6 to 8), secondary (grades 9 and 10), and higher-secondary levels (grades 11 and 12). There are some provisions for pre-primary education for children aged 3 to 5. However, this is available mostly in the cities and district headquarters and not considered to be a part of the formal education system (International Bureau of Education, 2011).

Nationwide public examination in the country includes the Primary Education Completion Examination (PECE) held at the end of five years of schooling, Junior Secondary Completion (JSC) at the completion of eighth grade, Secondary School Certificate (SSC) examination after ten years of schooling, and Higher Secondary Certificate (HSC) after 12 years. The enrolment age for primary education is six years and higher-secondary education is generally completed by age 17-18 (Pouezevara & Khan, 2007). After successfully completing HSC, students can enrol in higher education in universities, post-HSC level colleges and institutes offering professional, technical, technological and other types of education (International Bureau of Education, 2011).

Primary education is compulsory in Bangladesh. Statistics suggest that Bangladesh is currently running “one of the largest primary education systems in the world” (Habiba, Abedin, & Shaw, 2013; NAPE, 2018; UNICEF, 2009, p. 1). According to data presented in the Annual Primary School Census (ASPC) (Directorate of Primary Education [DPE], 2017a), there were more than 133,000 primary schools in

Bangladesh attended by over 17.25 million students. The number of teachers in these schools was above 500,000. Most of the schools at the primary level are government controlled (publicly financed and delivered) (Asadullah & Chaudhury, 2013; CAMPE, 2015). For all the government primary schools and for a vast majority of other schools, the language of instruction is Bangla (the official language of Bangladesh). However, some private enterprises offer education where the language of instruction is English. There are some schools which are known as “English Version”. These schools offer education based on the Bangladeshi curriculum but the instruction and teaching materials (including the textbooks) are all in English. There are also some schools commonly known as “English Medium” that offer education based on a recognised international curriculum (e.g., Edexcel, Cambridge International Examination, etc.). In addition, there is a Madrasa-based education system with emphasis on learning Arabic and religion-based education (Lutfuzzaman, Muhammad, & Hasan, 2006). The medium of instruction in the Madrasas (religious schools) is also Bangla.

The current curriculum of primary education in Bangladesh is competency-based. Competencies are defined as “the acquired knowledge, ability and viewpoint when these could be applied in real life at the right time” (Uddin, 2005, p. 5). The National Curriculum and Textbook Board (NCTB) of Bangladesh started implementing the competency-based curriculum and textbooks for all subjects after primary education was made compulsory (Roy, 2016) and this was gradually rolled out to all primary schools in the country in phases from grade 1 to grade 5 during 1992-1996 (NCTB, 2012). There were 53 terminal competencies in 1992 that students were supposed to achieve after completion of grade 5. These have been revised over the

years and now there are 29 terminal competencies in the latest curriculum (NCTB, 2012).

According to recent data, Bangladesh has made remarkable progress in improving some of the quantitative parameters (such as net enrolment rate, adjusted net enrolment ratio, gender parity, completion rate, etc.) of primary education (CAMPE, 2015; NAPE, 2018; Shekh, 2006). In addition, the public expenditure allocation for primary education as a portion of total public expenditure on education has gone up from 37% (in 2005) to 45% (in 2010) (Asadullah & Chaudhury, 2013). However, even with the quantitative gains and large investments, the learning achievement and quality of education is still considered to be a big concern and the education system is struggling to ensure an acceptable level of learning achievement for children completing primary education (CAMPE, 2015; Ministry of Primary and Mass Education [MoPME], 2015; NAPE, 2018; Zia-us-Sabur & Ahmed, 2010). Some researchers even argue that the quality of education has been deteriorating (e.g., Shekh, 2006).

Researchers (e.g., Islam, 2016; Richards & Islam, 2018) suggest that the achievements in quantitative parameters (i.e., gains in enrolment rate, completion rate, etc.) in Bangladesh's primary education were driven by international education milestones such as the Millennium Development Goals (MDG) and Education For All (EFA). International development partners like the World Bank also lend allusive support to such claims. For example, the World Development Report 2018 suggested that many governments and donors focused on and achieved the quantitative targets for the Millennium Development Goals, and now the Sustainable Development Goals

(SDGs) focus on better learning (World Bank, 2018b). Richards and Islam (2018) argue that Bangladesh is among the countries that fulfilled the quantitative requirements but lacks in quality of education. It is one of the countries that satisfied the “letter” of the international obligation (in terms of attaining quantitative goals) without heeding the “spirit” (i.e., attaining quality education). However, “schooling without learning is of little value” (Richards & Islam, 2018, p. 55).

Review of different publications regarding primary education in Bangladesh highlights how the quality of education is an issue. A report produced by the Bangladesh Bureau of Educational Information and Statistics and the Ministry of Women and Children Affairs (BANBEIS & MWCA, 2005) identified “lingering poor quality of primary education” (p. 7) as the biggest problem that Bangladesh is facing in the pursuit of its educational goals and commented that for most children the achievement and competency levels are very low. Chowdhury, Nath, Choudhury, and Ahmed (2002) mentioned that almost two-thirds of those who complete primary education have been found in their study to be functionally illiterate. Similarly, a report by Asian Development Bank (2011) mentioned that even with the rapid growth in enrolments and strong overall growth, the quality of primary education in Bangladesh is not at a satisfactory level yet. As such, improving the quality of primary education remains one of Bangladesh’s highest priorities (World Bank, 2011).

Different policy documents from the government of Bangladesh (e.g., the National Education Policy, published by the Ministry of Education, 2010) and international development partners (e.g., World Bank, 2013) recognise the importance of quality in education and place emphasis on providing quality education and creating an environment to improve quality. In recent publications, the Directorate of Primary

Education (DPE, 2018) highlighted improving educational quality as a top strategic priority for primary education in Bangladesh while World Bank (2016) suggested that improvements in the quality of education have significant social and economic impact and can help propel the country into higher productivity-led growth. Achievement of quantitative goals, such as increased enrolment while a crisis in quality of education unfolds, has been a point of concern in many parts of the world, especially in developing countries and significantly fewer children have been observed to achieve basic level of proficiency in Mathematics and reading in such situations (Sawada, Mahmud, Seki, Le, & Kawarazaki, 2017).

1.2 Mathematics in Bangladesh's primary curriculum

As mentioned earlier, Bangladesh has a competency-based curriculum for primary education. Similar to all the other subjects at the primary level, the mathematics curriculum is also competency-based. Mathematics is taught as a separate subject in all the different streams of primary education (Bangla medium, English medium, Madrasa-based, and others) and is a compulsory subject for all students in the mainstream education system up to the age of 16 (i.e., grade 10) (Lutfuzzaman et al., 2006). The current primary curriculum in Bangladesh was revised in 2011 based on the recommendations of the 2010 National Education Policy. It outlines the revised aims and objectives of primary education. There are 13 objectives of primary education in Bangladesh. Among them, two are especially relevant to mathematics, these are: "3. To acquire knowledge of the principles and processes of science and technology, use these in solving problems and help develop scientific attitude and be inquisitive."

(NCTB, 2012, p. 13), and “5. To help acquire the competencies of using-mathematical concepts, logical thinking and problem solving.” (NCTB, 2012, p. 13).

The curriculum also describes the terminal competencies students are supposed to achieve at the end of primary education (i.e., after completion of grade five). Currently there are 29 terminal competencies mentioned in the curriculum. These competencies are further broken down into subject-wise competencies (competencies for each subject taught at the primary level) and grade-wise competencies (competencies achievable at the end of each grade, a collection of all subject-wise competencies achievable for that specific grade) (Roy, 2016). There are three terminal competencies that are relevant for mathematics:

- 7. To develop the habit of solving problems and science mindedness through following the principles and processes of science and logical thinking,
- 11. To acquire the concepts and skills of Mathematics
- 12. To be able to solve Mathematical problems through logical thinking.

(NCTB, 2012, p. 14)

In addition to looking at how mathematics has been articulated in the primary curriculum in Bangladesh, assessment of mathematics was explored for the background information of this study. The current assessment practices in primary mathematics were one of the factors that formed the rationale for this research.

The Bangladeshi primary curriculum suggests that for a competency-based curriculum it is necessary to have provision for ongoing assessment of learners’

progress and provision for remedial teaching in order to address the deficiencies of learning. It also suggests that teachers should diagnose the learning deficiencies of the learners during their day-to-day teaching-learning activities and help students achieve mastery learning by providing said remedial measures. However, even though the importance of ongoing assessment and remedial teaching is mentioned in the curriculum, there is no detailed discussion regarding how to conduct these assessments for mathematics (or other subjects). In fact, there is very little guidance in the curriculum about how assessment should be conducted.

Personal experiences of the researcher and anecdotal evidence about the primary mathematics classrooms in Bangladesh suggest that there is a paucity of diverse assessment practices at the school level. However, it was not possible to support or disclaim these ideas on the basis of any empirical research as no study focusing on the assessment practices of mathematics classrooms at the primary level of Bangladeshi education could be identified. To have at least an inkling about what the theoretical situation (as opposed to actual practice) in the classroom is, teachers' guides (which are called "teacher editions" in Bangladesh) for the subject mathematics in grades 3 to 5 were reviewed. This review specifically focused on content related to arithmetic word problem solving or word problem solving in general because of their relevance to this study. These teacher editions are written as parallel to the mathematics textbooks students are given and each page of the student's book is represented in the teacher edition with explanations and advice about how to teach the content. It should be noted that though the revised curriculum and revised textbooks were implemented in 2012, the teacher editions were not published until August 2016 and were not readily available for teachers or

for review by researchers. The following sections briefly discuss assessment of word problem solving as described in the teacher editions. A more detailed representation of the learning and assessment activities mentioned against specific learning outcomes in the teacher editions is included in Appendix 1.1. All the teacher editions are in Bangla and the excerpts used here were translated for this discussion.

Teacher edition primary mathematics grade 3 (NCTB, 2016b)

The grade 3 textbook for students has a few separate chapters that deal with word problems for addition and subtraction operations. Review of the assessment instruction for teachers showed that they were advised to focus on whether students were able to successfully do the addition or subtraction, that is, it is considered as a correct/incorrect, able/unable assessment situation. Teachers are instructed to use observation, oral and written responses to determine whether students were able to use the operation and there is no mention of checking any other aspect of the word problem solving (e.g., whether the students had difficulty in understanding the word problem). There are word problems in later chapters of the book (e.g., currency conversion, measurement of length, weight and time, etc.) and the assessment instruction is similar to the previous chapters (i.e., focusing on the operation).

Teacher edition primary mathematics grade 4 (NCTB, 2016a)

There is not a lot of emphasis on word problem solving in the grade 4 mathematics book/teacher edition. Addition and subtraction are discussed in the second chapter of this book and chapter five discusses some word problems involving addition,

subtraction, multiplication and division. Even though there are some addition and subtraction word problems discussed in the second chapter, similar to grade 3, the focus of the teacher edition is on the operation and not the word problem solving. For example, while discussing word problems involving addition, assignment and observation are suggested as assessment methods to determine whether students were able to do addition of not more than five-digit numbers. Teachers are instructed to determine if the students were able to do the addition correctly. There is no mention of any assessment method/strategy to check if the students had difficulty in any aspect other than the operation. Similar instructions/suggestions are given for subtraction as well. However, though not within the assessment instructions, as part of the discussion on teaching-learning activities, there is a suggestion that teachers should be conscious that students are able to select the appropriate mathematical sentence based on the word problem solving item. There are some word problems discussed in later chapters of the book (contents on measurement, conversion of measurement units, time, etc.). Similar to the earlier chapters, assessment techniques like assignment, oral/written questions are mentioned as assessment techniques and the instruction is to focus on the content (e.g., conversion of units, time, etc.). There is no mention of checking if students were able to understand the problem or focusing on any other difficulties they might have, including understanding real life situations.

Teacher edition primary mathematics grade 5 (NCTB, 2016c)

The teacher edition for grade 5 has a chapter (Chapter 3) that covers mathematical problems involving the four operations (addition, subtraction, multiplication, and

division). Similar to the grade 3 and grade 4 teacher editions, the emphasis is on the operations, in this instance on how to use multiple operations together to solve a problem. Word problems are discussed but the focus is on the operations rather than on how to get to the operations from the word problems. Teachers are instructed to use various assessment techniques/methods such as observation, assignment, oral and written questions but the instructions predominantly focus on utilising parenthesis for solving equations.

As mentioned earlier, the teacher editions were reviewed as a proxy for understanding what type of assessments might be happening in the primary schools for Arithmetic Word Problem Solving. From this review, it is apparent that the focus of assessment is the operation (addition, subtraction, etc.) that the students would have to do as part of the AWPS and no other skill is likely to be assessed by the teachers. Even though the curriculum highlights the necessity of ongoing assessment and remedial teaching to address the deficiencies of learning, the prescribed assessments are not fine-grained enough to identify specific difficulties students might face and there is no indication of how assessment results could be used to address the learning deficiencies. Furthermore, it was noted that there is a paucity of empirical research on mathematics assessment at the primary level in Bangladesh.

1.3 Mathematics proficiency of primary students in Bangladesh

As previously highlighted, literature suggests that quality of education is a big concern for the Bangladeshi primary education sector. To understand the quality of learning in mathematics, this section reviews some assessment results and research findings

related to students' mathematics proficiency. Though some of the assessments discussed are of questionable quality themselves, the discourse is helpful for understanding the context of Bangladesh and situating the study. Besides, these details form part of the rationale for conducting this research.

There are a few sources of data that can be consulted to understand Bangladeshi primary students' mathematics proficiency. These can broadly be divided into three categories: Government administered large-scale assessment, NGO administered large-scale assessment and small-scale research studies.

1.3.1 Government administered large-scale assessment

There are two large scale tests administered by the government that can be utilised to understand students' performance in mathematics. The first of these is the Primary Education Completion Examination (PECE) and the second is the National Student Assessment (NSA). PECE is a high-stakes test that is administered every year with all grade 5 students and NSA is a large-sample based low-stakes test administered in alternative years.

1.3.1.1 Primary Education Completion Examination (PECE)

Since 2009, all mainstream primary students in Bangladesh must pass the Primary Education Completion Examination (PECE) at the end of grade 5 to move on to secondary education. There is a parallel examination called Ebtedayee Education Completion Examination (EECE) for students in the Madrasa stream, which was introduced in 2010. PECE is the largest public examination in Bangladesh where students attend centrally managed examinations for all grade 5 subjects. The different responsibilities related to writing tests, administering them and checking the answer

scripts are shared by the Directorate of Primary Education (DPE) and National Academy of Primary Education (NAPE) under the auspices of the Ministry of Primary and Mass Education (MoPME). The tests are administered nationwide at the same time with a common set of items and the results are published applying a grading system (Dey & Siddiquee, 2012). According to DPE data, more than 2.8 million students appeared in the PECE examinations each year from 2014 to 2016 (Directorate of Primary Education [DPE], 2017b).

PECE has been heavily criticised for having inflated scores and pass rates over the last few years. However, even with the supposed inflation of scores, students' performance in Mathematics has been poor compared to other subjects since the beginning of this assessment. The pass rate is very high in absolute numbers though. Hossain (2015) identified that the pass rate for mathematics was the lowest among all the PECE subjects from 2009 to 2012 (2009 – 90.57%, 2010 – 95.58%, 2011 – 98.19%, 2012 – 97.87%). Pass rates in mathematics for the subsequent years were also lower than for the other subjects. From 2013 to 2015 Mathematics had the second lowest pass rate (2013 – 99.36%, 2014 – 98.91%, 2015 – 99.36%) with English having the lowest pass rate with a very small margin (2013 – 99.10%, 2014 – 98.89%, 2015 – 99.18%) (Directorate of Primary Education [DPE], 2015, 2016; Ministry of Primary and Mass Education [MoPME], 2014). It is noticeable how more than 90% of students passed in the Mathematics tests in all these years. But researchers do not consider PECE results as a very dependable indicator of students' achievement in general and have identified a number of issues with them. This includes the low threshold for passing (students only have to achieve 33% score for a pass), lax grading practices/ grade inflation, test-question leak (questions distributed on social media, email or via

text messages on phone days or hours before the tests), teachers writing on students' answer scripts, proxy test-takers, and so on (CAMPE, 2014; Opel, Zaman, Khanom, & Aboud, 2012; Richards & Islam, 2018). Such issues make the PECE an unreliable measure for student performance (Richards & Vining, 2016).

1.3.1.2 National Student Assessment (NSA)

The National Student Assessment (NSA) is the only large-scale sample-based assessment used in by the Bangladeshi government to assess the performance of the primary education system. It was “initiated in 2006 by the Ministry of Primary and Mass Education (MoPME) to assess achievement in primary education” (DPE, 2018, p. X). The intention was to collect critical information about the effectiveness of the second Primary Education Development Program (PEDP II) interventions in improving the quality of education and students' learning outcomes (Hossain, 2015). PEDP is a sector-wide program for the improvement of primary education in Bangladesh, currently in its fourth phase (i.e., PEDP IV). NSA is managed by the Monitoring and Evaluation Division of the Directorate of Primary Education (DPE) under MoPME with technical support from external research institutes (in 2011 and 2013 from the Australian Council for Educational Research – ACER and in 2015 and 2017 from the American Institutes for Research – AIR).

Since 2006, NSA has been conducted every other year (with the exception of the 2010 assessment being pushed back to 2011). It is a sample-based assessment that focuses on understanding the primary education sector's condition by testing a representative sample of grade 3 and grade 5 students from the whole country. NSA

tests are “conducted on the two foundational subjects, Bangla and Mathematics and provide a method for tracking student learning outcomes at specific points in the primary cycle” (Cambridge Education, 2016, p. 33). According to DPE (2018) the 2017 iteration of NSA employed a stratified sampling method to test 28000 grade 3 and 24000 grade 5 students from 1600 schools and education centres. The tests were constructed based on the competencies outlined in the 2012 national curriculum and the items were a mix of selected response and constructed response items. This is a distinct trait of the items used in NSA, that is, the items in this assessment focus on skills and competencies outlined in the national primary curriculum in lieu of focusing on recall of the textbook content (CAMPE, 2014; Islam, 2016; Roy, 2016). Another crucial difference between the NSA and the PECE is its use of Item Response Theory (IRT). Whereas PECE results are raw score based, a common measurement scale based on IRT (with a range of 60 to 180) is constructed for interpreting the NSA scores. Results from this assessment are communicated through performance bands, where “Band 1 (students working well below primary 3 level) is considered the basic level of proficiency while Band 5 (students working at primary 5 level) is considered the highest skill level” (Cambridge Education, 2016, p. 33).

Students’ performance in the NSA Mathematics tests over the last few years has shown poor achievement in the subject and has always been lower compared to Bangla scores in the same assessment. According to NSA 2011 results, half (50%) of grade 3 students reached grade level expectations (i.e., mastered grade 3 level competencies) whereas only one-third (33%) of the grade 5 students reached their grade level expectations (i.e., mastered grade 5 level competencies) (Ministry of Primary and Mass Education [MoPME], 2015). If we compare this data with the PECE

results from the same year, 98.11% students passed the mathematics test in PECE in 2011 (Hossain, 2015).

DPE (2018) outlines the low achievement of students in the NSA mathematics tests from 2013 to 2017. The report (DPE, 2018) shows that student performance has declined in grade 3 mathematics with 57% of students reaching grade level expectations in 2013 and 41% reaching the same in both 2015 and 2017. Similarly, in grade 5 mathematics 25% of students reached grade level expectations in 2013 (compared to 33% in 2011) whereas 10% of students in 2015 and 17% of students in 2017 reached the same level. If we compare this with grade 5 students' mathematics results in the PECE, in both 2013 and 2015 99.36% of students passed the mathematics test. It is likely that many students are passing the PECE mathematics tests achieving only a small portion of the grade level competencies for the subject.

In addition to comparing with the 2013 results, DPE (2018) revised the performance levels for the 2015 and 2017 NSA tests for some additional analysis. However, even with the revised performance levels for these two years, only 28% of grade 3 students in 2015 and 34% in 2017 were performing at the proficient or advanced level in mathematics. Likewise, only 25% of grade 5 students in 2015 and 32% in 2017 were at proficient or advanced level in the revised interpretation. Overall, data from the NSA test results suggest that a large proportion of the students were performing below their expected grade levels in Mathematics. Furthermore, Islam (2016) argues that even the low achievement shown by students in the NSA might be an overestimate of their abilities, as there are suggestions that the

assessment may be at too low a level of performance. For example, comparing the literacy items from NSA with Progress for International Reading Literacy Study (PIRLS) items, a World Bank report (World Bank, 2013) found that the NSA literacy items were relatively easier to master.

1.3.2 NGO administered large scale assessment

The Campaign for Popular Education (CAMPE) administers the only large-scale student assessment in the primary education sector of Bangladesh outside of the government assessments that periodically provides national data on learning achievements (Hossain, 2015). CAMPE, which is an advocacy and campaign network of approximately 1000 non-government organisations (NGOs), formed a civil society group called Education Watch in 1998 as a platform for educational development discourse. Although the competency-based curriculum and textbooks were introduced in Bangladesh from 1992, the initiative for the first formal competency-based large-scale assessment was taken by the Education Watch (EW) group only in 2000 (Roy, 2016). While designing the assessment instrument, the EW group selected 27 competencies from the 53 competencies in operation in 2000. It covers five competencies of Mathematics and has 15 items on this subject. The same test has been administered in 2000, 2008 and 2014 (CAMPE, 2015). Though the national terminal competencies were revised after construction of the instrument, the EW instrument has not been changed as the competencies it measures are still considered to be part of the revised set of competencies. One reason for this being a viable option is that, whilst the national terminal competencies were reduced to 29 from 53, in many cases the old competencies were rephrased and merged to form the new

competencies. Education Watch Report 2015 (CAMPE, 2015) describes the findings from the most recent competency-based test taken by grade 5 students in 2014. This iteration of the assessment included 5375 students from 309 educational institutions from all over Bangladesh. In 2014, students on average achieved 74.5% of the competencies in the test. They achieved 69.2% of the mathematics competencies tested, which is the second lowest after English. Comparing this result with previous administration of the assessment, the report highlights that achievement of Mathematics competency has always been the second lowest with students achieving 48% competencies in 2000 and 61.2% in 2008.

It is interesting to note that although students' achievement in Mathematics remains one of the lowest, similar to PECE results, there is a trend of improvement over the years.

1.3.3 Small-scale research studies

There are not many empirical data-based research studies that deal with primary mathematics education or primary students' mathematics achievement in Bangladesh. This section summarises the few studies that give us some insight about primary mathematics education in Bangladesh.

Richards and Islam (2018) conducted a survey in the northern rural districts of Bangladesh where they assessed students' "ability to read and solve mathematical problems" (p. 56) at the grade 2 level of the Bangladeshi curriculum. Their sample included 555 students studying in grades 1-5 (aged 6 – 16) in 18 different schools. Based on their data the authors suggested that students' performance was weaker in

Mathematics compared to reading and only 18% of the grade 3 students were found to be working at grade 3 level.

NAPE (2015) conducted a small-scale research study to determine the suitability of competency-based items in the 2014 PECE examination. Whilst this was done in order to understand how the competency-based items compared with other items, findings from this study give us some idea about students' performance in mathematical competencies. The report showed that the average score of students in the competency-based items was 21.94 (63%). Furthermore, seven constructed response items were included as part of this study that required students to do mathematical problem solving with the application of the four operations (addition, subtraction, multiplication, and division). The mean score for these items was 10 out of a possible 20 (i.e., 50% correct). The report suggested that the reason for this poor performance in mathematical problem solving was because the students were not accustomed to solving novel problems by themselves and were more used to responding to textbook items.

In his research, Hossain (2015) used terminal examination (final examination to pass a grade level) and PECE examination data of grade 3 to grade 5 students from 16 sample schools to determine the learning achievement situation in mathematics. These schools were located in four different divisions of Bangladesh. He found that for these three grades in the sample schools, if a student failed a subject, it was Mathematics in most cases. Among the students who failed, 88.9% in grade 3, 79.65% in grade 4 and 77.7% in grade 5 failed in Mathematics.

Asadullah and Chaudhury (2013) conducted a small-scale study in rural Bangladesh where they tested mathematics achievement of ten to 18-year-old children using a rudimentary test that was designed based on primary level mathematics competencies. Their assessment instrument had both oral and written mathematics components and was administered in households to include students from different backgrounds (e.g., primary school graduates who went on to study in the secondary level, dropouts, etc.). The authors found that even with the very easy competence test, mathematics achievement of children was very low, with only 43% attaining basic competence in oral assessment and 52% attaining basic competence in written mathematics. In addition, they observed that 52% of the children who had successfully completed primary education failed the written competency test and 62% failed the oral test. Furthermore, their data showed that students needed additional years of schooling beyond the primary level to attain basic competency in mathematics.

In his research, Uddin (2005) reviewed two survey results and concluded that pupils' performances in mathematics competencies were unsatisfactory in general and they were especially weak in the following areas: "solving problems in real life, the four fundamental rules and measurement" (Uddin, 2005, p. 10).

Overall, these few empirical evidence-based studies available on mathematics achievement/performance of students at the primary level in Bangladesh show that student performance is below-par and there is a need for more studies to be conducted in this area.

1.4 Primary teachers in Bangladesh

Though teacher quality is considered to be an important factor for influencing students' learning, this poses a challenge in Bangladesh and various reports have identified the quality of teaching in the country to be unsatisfactory. For example, UNICEF (2009) described the quality of the teaching learning processes, poor qualifications and lack of motivation of teachers as major challenges for primary education in Bangladesh. In addition, a report by UNESCO (2006) suggested that Bangladesh is among the lowest achieving countries in terms of meeting the minimum standard for teaching. Only 37% of the teachers met the required qualifications in junior secondary level (third from the bottom in the list) and 21% of teachers in secondary level fulfilled the qualifications (lowest for any country). One reason for poor teacher quality could be lack of formal training as UNICEF (2009) reported that about 24% of teachers in government primary schools and registered non-government primary schools were untrained.

Shohel and Banks (2010) also highlighted teacher education and training in Bangladesh to be a major reason for the country's low success rate in educating its citizens. Other researchers have highlighted primary and secondary teachers' lack of understanding in selecting and constructing assessment techniques suitable for instructional decisions to be another point of concern, as the teachers still use tests or simple question-answer methods for most of their assessment tasks (Ahsan, 2009; Khan, Azim, & Ahmed, 2009). Although little research is available on primary teachers' use of assessment, primary schools were generally found to be focusing on basic memorising facts/ rote memorisation (Opel et al., 2012; UNICEF, 2009).

1.5 Rationale of this study

As mentioned above, Bangladesh is struggling in achieving quality primary education despite many efforts. It has been pointed out by various reports and research studies numerous times that Bangladesh has previously focused on achieving quantity in primary education and now must focus on quality. This pronounced lack of quality is particularly true in the case of how students perform in Mathematics. Both the Primary Education Completion Examination (PECE) and the representative sample based National Student Assessment (NSA) demonstrate students' low performance in Mathematics (the PECE gives a relative view whereas NSA gives an absolute view). Results from PECE over the last few years show that mathematics achievement remains one of the lowest for students, with the mathematics pass rate being the lowest from 2009 to 2012 (Hossain, 2015) and being the second lowest from 2013 to 2015. Despite the high percentage of students passing (i.e., getting 33% of the score) in PECE mathematics based on the raw score data, their performance is relatively poor when compared to results in other subjects. On the other hand, the poor performance students demonstrated in the NSA mathematics can be regarded as a more robust indication of their mathematics achievement since this is calculated using IRT. In the 2017 NSA, only 41% of grade 3 students and 17% of grade 5 students achieved their grade level expectations in mathematics (DPE, 2018). Such poor performance in this subject has been the norm in previous iterations of the NSA too.

Anecdotal evidence also supports the idea that students need to understand Mathematics better. For example, Secondary School teachers often complain that students' mathematical knowledge built during the primary years is very weak, which

causes continuous weakness and fear in students in relation to learning Mathematics (Institute of Educational Development [IED], 2009). Hossain (2015) commented that the lack of quality teaching and learning of primary mathematics creates low achievers which can increase drop-out rates and makes a learner disinterested in learning science, which then consequently acts as a bar to achieving science-based objectives and aims of primary education.

Education systems worldwide have been known to utilise assessment practices to induce change in educational practices and it can be argued that changing assessment practice might help in developing the quality of primary education in Bangladesh. Since the commencement of formal education, assessment has always played a role and has progressively established itself as an integral part of teaching and learning. Nowadays assessment is utilised for a number of educational purposes including prediction, selection, placement, diagnosis, evaluation, grading, guidance, and administration (Lamprianou & Ahanasou, 2009). Zhang and Burry-Stock (2003) suggest that assessment works as an apparatus for cultivating classroom learning and it is possible to improve classroom learning by improving teachers' assessment practices and skills. Due to the crucial role appropriate classroom assessment practices play in maximising student attainment and motivation, they have been acknowledged as one of the most important components of the teaching-learning process (Lukin, Bandalos, Eckout, & Mickelson, 2004; Nolen, 2011).

The national curriculum of Bangladesh suggests that ongoing assessment of learners' progress and remedial teaching to remove the deficiencies of learning are required in a competency-based curriculum (NCTB, 2012). While the curriculum and

textbooks are competency based, researchers have commented that general assessment procedures in Bangladesh are mainly content-based where central and organised competency-based assessments are rare (Banu, 2009; Ministry of Primary and Mass Education [MoPME], 2015; Roy, 2016). Similarly, school assessments also focus on memorising facts/contents (UNICEF, 2009). However, as mentioned earlier, changes in assessment in the right direction can play a major role in improving the quality of education. For example, a World Bank report suggested “prioritizing high quality learning in lower grades and building strong foundational cognitive and non-cognitive skills” (World Bank, 2013, p. 69). This report emphasised “changing what is valued in the examination system (focusing more on learning competencies)” (p. 69) as an initiative for improving students’ learning outcomes. This recommendation is echoed by Richards and Islam (2018) who argue that one prerequisite for achieving quality (in Bangladesh primary education) is to assess learning outcomes adequately through an appropriate tool.

The general term ‘assessment’ is defined by Black and Wiliam (2010) as the “activities undertaken by teachers – and by their students in assessing themselves – that provide information to be used as feedback to modify teaching and learning activities.”(p. 82). These activities can be considered as formative assessment if the evidence gathered is utilised to tailor teaching according to the students’ needs and help them learn more effectively (Black & Wiliam, 2010; Dixon & Worrell, 2016). The formative assessments teachers conduct in their classrooms help them in understanding students’ learning progression and inform their teaching. The power of using such assessment resides in providing information and interpretations about the discrepancy between students’ current learning progression and the learning goals,

which can inform the teacher about how to choose effective teaching strategies (Hattie & Timperley, 2007). On the other hand, summative assessment is an assessment of how much learning has taken place (Gardner, 2010) and captures “evidence of what has been learned at certain times” (Dolin, Black, Harlen, & Tiberghien, 2018, p. 54). Such assessments are often cumulative, record what a student has learned, the quality of their learning and sometimes compare their learning performance against some predetermined standards (National Research Council, 2001). However, simply using summative assessment to identify the level of proficiencies of students cannot effectively assist in students’ learning process (Black & Wiliam, 1998; Crooks, 1988). Formative assessment in classroom teaching and learning can be used as a tool to provide constructive/diagnostic information about students’ learning process. In order to use formative assessment to inform the effectiveness of their teaching strategies, teachers need to devise activities and questions that can enable the gathering of adequate information so they can provide feedback to students based on their needs (Hattie & Timperley, 2007).

The current classroom assessment and formative assessment practices of teachers in Bangladesh are not at a satisfactory level. Rahman (2010) suggests that, in Bangladeshi teachers’ classroom assessments, assessment of higher order learning is almost always ignored. Teachers ask questions from the textbooks and often the phrasing of the questions encourages students to answer by recalling simple facts. In view of the current practices in Bangladesh, Ahsan (2009) called for classroom assessment practices that incorporate effective and multiple ways of assessing students. Rahman (2010) further noted how teachers’ administrations of formative assessment suffer as they fail to allocate sufficient quiet time for students to think and

prepare an answer. Teachers' decision making based on assessment results is limited to asking the poor-performing students to study harder in most cases or might result in a parent-teacher meeting if the situation is extreme.

This research focuses on developing an assessment that will make it easier for teachers to improve their assessment practices (both formative and summative) by identifying areas where students are facing difficulty. However, this study focuses on one specific area of mathematics, namely arithmetic word problem solving. Review of existing literature contributed to narrowing down the area of research to this specific focus. This will be further discussed in the next chapter.

As mentioned earlier, there are objectives and competencies related to mathematical problem solving within the curriculum of mathematics which highlight the importance given to this area in primary education in Bangladesh. Considering the emphasis attributed to problem solving worldwide, this is an area that needs to be well-developed in the primary graduates. However, as previously discussed, little research is available on Bangladeshi primary mathematics education in general and no studies were found that focused specifically on arithmetic word problem solving. This study endeavours to contribute to filling this gap in the field of research.

1.6 Research objectives and research questions

This research focuses on arithmetic word problem solving and aims to develop exemplary assessment materials that will enable teachers to identify specific difficulties students face while solving arithmetic word problems. To achieve this aim, the research objectives are to:

- (a) develop a theoretical framework for arithmetic word problem solving using an iterative process (as discussed by Wilson, 2005)
- (b) construct tasks aligned with the framework,
- (c) try them out in both small- and large-scale studies,
- (d) investigate the results using qualitative and quantitative methods, and
- (e) evaluate the success of the process by investigating the validity and reliability of the results.

Based on these objectives, the research questions for this study are:

- i. What would be a theoretical framework for arithmetic word problem solving?
- ii. What items can be used to assess various stages/steps of this theoretical framework?
- iii. How can this framework be used to identify students' difficulty in solving arithmetic word problem?
- iv. How valid and reliable is this process of identifying students' difficulty?

1.7 Structure of the thesis

This thesis is divided into seven chapters. Chapter 1, the current chapter, discusses the educational context of Bangladesh, outlines the rationale for the study and delineates the research objectives and research questions of this study. First, the overall structure of the education system in Bangladesh is briefly discussed. Then a more in-

depth discussion of the primary level is presented. The next section describes how mathematical problem solving is articulated in the Bangladeshi primary curriculum. In addition, discussions about assessment of mathematical problem solving in the teachers' support materials (teacher editions) are outlined. In the subsequent sections, some assessment results and research findings related to students' mathematical proficiency are reviewed to determine the quality of learning in mathematics. Next the quality of primary teachers in Bangladesh is briefly discussed. The subsequent section presents the rationale for this study based on the previous discussions and then the research objectives and research questions of the study are presented.

Chapter 2 presents the review of literature where a broad to narrow approach has been taken. First, literature related to definition of the problem, problem solving, historical background of the concepts of problem solving, and domain specific problem solving are discussed. Then literature on mathematical problem solving, teaching of mathematical problem solving, and arithmetic problem solving are reviewed. Subsequently, based on the previous discussions, a theoretical framework for assessment of arithmetic word problem solving is hypothesised. Finally, literature on validity and reliability is discussed in order to define the approaches taken in this study to establish the validity and reliability of the hypothesised assessment framework.

Chapter 3 describes the methodology and methods used in this study. First, different research philosophies are reviewed, and the research philosophy adopted in this study is outlined. Then the different data collection instruments used in this

research (e.g., students' test, students' cognitive interview, expert opinionnaire, etc.) and how they were developed is discussed. This section also outlines how the Arithmetic Word Problem Solving (AWPS) test used in this study was developed based on the hypothesised assessment framework. Subsequent sections report on the selection of participants for this study and the administration of the data collection instruments. The next sections delineate how the data were managed and analysed. Finally, the ethical considerations of this research are mentioned.

Data collection and analysis in this research was conducted in two phases. The first phase, which is called 'the pilot study' in this thesis, was conducted with a small sample of students. Students' tests were administered (with 62 students) and few cognitive interviews (6) were conducted. In addition, students' school examination scores for mathematics and Bangla were collected. Data collected in the pilot study helped in revising the test for the next phase. The second phase of the research, which is dubbed 'the final study' in this thesis was conducted with a larger sample and had multiple data sources. In addition to the students' tests, cognitive interviews, and school-test scores, expert opinions, and other data were collected in this phase.

Chapter 4 of this thesis presents the data analysis of the pilot study and the results. Item analysis of the pilot test (Wright map, item fit, response fit, etc.) is discussed in this chapter. In addition to the psychometric analysis, analysis of students' interviews and analysis of validity evidence based on relation to other variables (i.e., comparison of students' AWPS test score with their text-book item score and with their mathematics and Bangla scores from the school) are discussed.

This chapter concludes with a discussion on how the test items were revised based on the pilot data.

Chapter 5 discusses the analysis of quantitative data from the final phase of data collection. This chapter includes item analysis (Wright map, item fit, response fit, etc.), model fit analysis, and analysis of test reliability. In addition, analyses of validity evidence based on relation with other variables (e.g., comparison of AWPS test and textbook item scores, comparison of AWPS test and school test scores) are also discussed here. The final section of this chapter describes the analysis of validity evidence based on test content, that is, the analysis of expert opinionnaire data.

As part of gathering validity evidence, qualitative data were collected through students' cognitive interviews in the final phase of this study. Chapter 6 reports on the analysis of these qualitative data. Here, students' response processes for items in the different strands of the hypothesised assessment framework for arithmetic word problem solving are analysed. The chapter is organised based on the strands of the framework.

Chapter 7 summarises the findings of the study by reflecting on what the objectives of this study were and how they were fulfilled. The successes and limitations of this study, what could be done differently in future studies, what changes could be brought to the methodology of this study, and what other types of analysis could be conducted – are also discussed. Finally, the implications of this research for Bangladesh are discussed.

Chapter 2: Review of literature

In identifying the area of research for this study, an approach was taken where a broad concept was explored first and gradually narrowed down to a researchable area based on extant literature. The initial exploration of literature began with understanding what the problem is and why it is important. Before discussing the main variables of this study, the next sections present the theoretical background for this work discussing the definition of the problem, problem solving, historical background of concepts of problem solving, and domain specific problem solving. Then literature on mathematical problem solving, teaching of mathematical problem solving, and arithmetic problem solving are reviewed. Next, a theoretical framework for assessment of arithmetic word problem solving is hypothesised. Finally, literature on validity and reliability is discussed to determine the suitable approaches for investigating the validity and reliability of the hypothesised assessment framework.

2.1 What is a Problem?

According to Duncker (1945) “a problem arises when a living creature has a goal but does not know how this goal is to be reached” (p. 1). Mayer and Wittrock (2006) reiterate this definition and further explain that “a problem occurs when a problem solver wants to transform a problem situation from the given state to the goal state but lacks an obvious method for accomplishing the transformation” (p. 288). Charles and Lester (1982) broke it down to ‘task’ level rather than the goal or goal state and defined a problem to be a task that a person is facing and must find a solution for.

However, there is no readily available procedure to find a solution. Therefore, the person must make an attempt to find a solution. It is noticeable how in this definition they have included an actively solution seeking component. Blum and Niss (1991) took a different approach by considering 'problem' in terms of intellectually challenging questions. They defined a problem to be a situation that carries certain open questions and intellectually challenges individuals who do not have the direct methods/ procedures/ algorithms at the moment to answer the questions. In this definition, problem is more person-dependent, since the same situation might not be equally challenging for every person. Although such definitions have emerged that focus on different aspects of a problem, Duncker's (1945) depiction of problem with an initial state and a goal state has been the one widely adapted by different authors and studies over the last few years (e.g., Jonassen, 2000; Mayer, 2013; Newell & Simon, 1972; OECD, 2012; Wu, 2003).

2.2 Problem Solving

Most definitions of problem solving is based on Duncker's (1945) definition of problem, that is, they define problem solving as a process of reaching a goal state from an initial state when no method of reaching the goal is readily available to the problem solver. For example, Anderson (1990) defined problem solving as "any goal directed sequence of cognitive operations" (p. 257). Other studies have included the absence of a readily available solution method as a factor and defined problem solving to be the "cognitive processing directed at achieving a goal when no solution method is obvious to the problem solver" (Mayer & Wittrock, 2006, p. 287). Robertson (2001),

Mayer (2013), and OECD (2014) also articulated a similar notion of problem solving. Simply put, problem solving is the attempt at achieving some outcome when there is no known method (for the individual trying to solve the problem) to achieve it (Schoenfeld, 2013). Reaching the goal state might not be possible in a single step. Anderson (2015) considered this in his recent book and defined problem solving as “goal-directed behaviour that often involves setting sub-goals to enable the application of operators” (p. 183).

While this research focus on the cognitive processes and achieving the goal state, Lester and Kehle (2003) additionally mention non-cognitive factors and depict solving problem as resolving the “tension of ambiguity” (p. 510). They suggest that problem solving entails an individual engaging in “a variety of cognitive actions, each of which requires some knowledge and skill, and some of which are not routine. Furthermore, these cognitive actions are influenced by a number of non-cognitive factors” (p. 510). According to them, a number of factors (some of these are non-cognitive) need to be coordinated for successful problem solving. These include “previous experiences, knowledge, familiar representations and patterns of inference, and intuition” (p. 510). Problem solving then leads to “new representations and related patterns of inference that resolve the tension of ambiguity which prompted the original problem solving activity” (p. 510). It is noticeable how they mention intuition as a factor in successful problem solving. Kaur and Toh (2011) also include intuition while defining problem solving. They posit problem solving as a complex process requiring coordination of existing knowledge, experiences, understandings,

and intuitions. Successful coordination results in mitigating the challenges of a novel situation.

2.3 Historical background of concepts of problem solving

Psychologists have researched problem solving approaches and methods for more than a hundred years. The early conceptions of problem solving research were developed in Leipzig, Germany where “Wilhelm Wundt opened the world’s first psychology laboratory” (OECD, 2010, p. 35) in 1879. Though Wundt had initially rejected the idea of researching complex cognitive processing through experimental methods, a group of his former students based in Wurzburg began researching word association problems using an introspection technique and although they did not produce a new theoretical approach to problem solving, they successfully collected empirical evidence on mental processes (e.g., imageless thought) (Mayer, 2013). In the early 1900s, while researching with animals, Thorndike (1911) developed a theory of learning - the ‘law of effect’. This theory suggests that behaviours that are followed by positive reinforcements/ rewards are likely to be repeated. In one of his experiments, Thorndike put a cat inside a puzzle box. The cat could only escape if it hit a lever inside the box. On the first day of the experiment it took the cat some time to try different things before it pushed the lever and escaped the box. As the experiment went on for several days, the time needed for the cat to push the lever decreased suggesting the cat was learning how to escape, that is, associating pushing the lever with escaping. This theory of associationism is considered to be “the first major theoretical approach to take hold in the scientific study of problem solving” (Mayer,

2013, p. 771). Based on Thorndike's theory, behavioural psychologists believed that "problem solving was a process of trial and error" (Bahar, 2013, p. 17).

Study of problem solving gathered motion through the works of Gestalt psychologists in Germany during the 1930s. In contrast to behavioural psychologists, Gestalt psychologists believed that instead of individual associations, problem solving involves constructing a coherent mental structure. Problem solvers build internal representations of a problem while trying to find its solution. This usually leads to a flash of insight and helps them to identify the solution. For example, in a study of apes, Kohler (1925) experimented with an ape in the play yard where there were several empty crates and a banana overhead, out of reach. He noted that the ape did not try a number of different things before one worked (as suggested by behaviourists). It looked at the banana and the crates for a while and then with a flash of insight made a ladder with the crates and reached the banana. In a sudden visual reorganisation, the elements constructed themselves in a way that solved the problem. Insight ensues as the problem solver "suddenly moves from a state of not knowing how to solve a problem to a state of knowing how to solve it" (Mayer, 1995, p. 3). Cognitive psychologists and computational neuro-scientists criticised Gestalt theories for being descriptive instead of explanatory in nature (Bahar, 2013).

In research based on Gestalt theory, problem solving was mostly confined to laboratory experiments with simple tasks and it was believed that results from such tasks could be generalised to more complex problems. As researchers started to understand it was not possible to do so, problem solving moved beyond laboratory experiments and Gestalt problem solving made way for a new field of study –

cognitive science (Lester & Kehle, 2003). Newell and Simon (1972) were among the first to strengthen this line of research in human problem solving.

In discussing human problem solving, Newell and Simon (1972) considered humans as “Information Processing Systems (IPS)” and identified three crucial elements of problem solving: environments, problem spaces, and programs. They argued that human problem solving is to be understood through description of the task environment in which it takes place; the space that is used by the problem solver to represent the environment, the task, and the knowledge about it that s/he gradually accrues; and the program s/he assembles in order to approach the task. Problem solving in this approach consists of a problem space and a search heuristic. The problem space is a mental representation of the given state of the problem, the goal state, and all possible intermediate states. Search heuristics are ways of manoeuvring through the problem space to reach the goal state. This approach makes it possible to describe the problem solving very clearly (Mayer, 2013) and has become a major problem solving approach over the last few decades (Anderson, 2015).

Charles and Lester (1982) suggest that from 1930 to 1970 the prevailing discussions of problem solving had three assumptions: i) in theory, the goal was understanding the cognitive processes of the problem solver, ii) internal goals guided cognitive processes, and iii) all types of problems had essentially the same cognitive processes. Problems were treated as considerably more domain-general in order to tend to these assumptions. However, beginning in the 1970s, researchers gradually moved away from simple laboratory tasks and towards domain specific problem

solving (Charles & Lester, 1982). The next section will highlight the importance of problem solving being domain-specific.

2.4 Domain specific problem solving

Research on problem solving expertise suggests that a theory of problem solving should not be separated from the domain in which the problem is being solved (Tuma & Reif, 1980). When solving problems in some domains, domain-specific problem solving skills might need particular analytical skills for that domain in addition to complex general mental processes not bound to a specific domain (Greiff, Holt, & Funke, 2013). Sweller (1988) argues that when it comes to problem solving, the primary factor distinguishing experts from novices is the domain-specific knowledge in the form of schemas. Furthermore, research suggest that long-term memory, which is a crucial element of human cognition, does not store random, detached facts. Instead, large complexes of closely integrated information are stored together resulting in knowledge domain-specific problem solving competence, and an experienced problem solver of a specific domain will have acquired a huge number of domain-specific problem solving schemas in the long term memory (Sweller, Clark, & Kirschner, 2010).

While general problem solving has its own applications, strategies of general problem solving are often regarded as weak, even though widely applicable whereas domain-specific methods are considered to be more powerful despite being narrow-focused (Stacey, 2005). In order to gather an in-depth understanding about problem solving, it might be necessary to consider domain-specific problem solving.

Furthermore, Lester and Kehle (2003) identify that “the processes underlying complex problem solving in one domain were different from those underlying another domain” (p. 502). This can be another argument for researching domain-specific problem solving instead of general problem solving.

It is also important to focus on domain-specific problem solving when it comes to teaching problem solving. Mayer and Wittrock (2009) argued that there is enough cognitive science research to highlight how teaching problem solving within the context of specific subject domain would be effective. In lieu of teaching within a general course on how to solve problems, successful programs integrate problem solving within the exact context in which they are going to be used (Nickerson, 1999). Besides, it is easier to make progress on a specific domain rather than all problem solving (Mayer, 1982). In a review of research conducted in the recent past, Mayer and Wittrock (2006) identified three principles of problem solving, one of which emphasises that instead of attempting to develop general problem-solving heuristics (in students) , it is better to develop problem solving skills within specific domains.

This study does not focus on general problem solving. Instead, it is domain-specific and considers Mathematics as the domain of study, that is, the focus is on mathematical problem solving. Choice of mathematics as the content area for studying problem solving is common among both educational and psychological research (Wu, 2003). Sweller et al. (2010) highlights the importance of domain-specific mathematical problem. They argue that mathematical problem solving skills are “acquired through a large number of specific mathematical problem-solving strategies” (p. 1304) and future mathematicians can be effective problem solvers only

if they are provided with a substantial set of domain-specific schemas since “there are no separate, general problem solving strategies that can be learned” (p. 1304).

According to previous researchers, research on problem solving can be grouped into two categories i) research on how to teach problem solving and ii) research on how we think when we solve problems (Codina, Cañadas, & Castro, 2015). This particular study positions itself within the second group.

2.5 Mathematical problem solving

Problem solving is considered to be a fundamental goal of teaching mathematics. Mathematical problem solving plays a crucial role in the teaching and learning of mathematics. Researchers consider it as “a vehicle for the teaching and reinforcing of mathematical knowledge” (Tambara, 2015, p. 61), solving real life problems, and helping to meet everyday challenges (National Council of Teachers of Mathematics [NCTM], 2000; Stacey, 2005). “Solving problems is not only a goal of learning mathematics but also a major means of doing so” (National Council of Teachers of Mathematics [NCTM], 2000, p.52). That is, problem solving is also regarded as a key approach in learning mathematical concepts (Tai & Lin, 2015). A huge amount of teaching time is allocated for mathematical problems in the classroom (Hossain, 2015). Furthermore, it is a skill that can augment logical reasoning competence (Kaur & Toh, 2011). Some researchers (e.g., Halmos, 1980) even argue that problem solving can be considered as “the heart of mathematics” (p. 519) due to the pervasive role it plays.

Schroeder and Lester (1989) described three distinctive roles and purposes of problem solving which give a comprehensive notion regarding why it plays a crucial role in mathematics teaching-learning:

- Teaching *for* problem solving (teaching mathematical content for later use in solving mathematical problems)
 - Teaching *about* problem solving (teaching heuristic strategies to improve generic ability to solve problems)
 - Teaching *through* problem solving (teaching standard mathematical content by presenting non-routine problems involving this content)
- (as cited in Stacey, 2005, p. 345).

In addition to researchers, governments and policymakers of a number of countries have emphasised mathematical problem solving over the last few decades. If we consider recent mainstream approaches in Mathematics teaching within the major English-speaking countries such as UK, USA and Australia, we will see how they have emphasised problem solving with certain commonalities and explicit influences visible within them.

In the UK, from 1940 onwards, design of mathematics curricula became more open and innovators began to highlight problem solving as a vital part of doing mathematics (Burkhardt & Bell, 2007). In 1982 the Cockcroft Report advised that “mathematics teaching at all levels should include opportunities for problem solving” (Cockcroft Report, 1982, paragraph 249). One of the aims of the latest national curriculum for mathematics in England is to make sure that all students, “can solve problems by applying their mathematics to a variety of routine and non-routine

problems with increasing sophistication, including breaking down problems into series of simpler steps and preserving in seeking solutions” (Department of Education, 2014, para 2, bullet point 3).

In addition, problem solving is repeatedly mentioned and emphasised in the programs of study for different years and key stages.

In the US, the “Principles and Standards for School Mathematics” published by the National Council of Teachers of Mathematics (NCTM) in 1989 emphasised that problem solving should be the central focus of mathematics curriculum (National Council of Teachers of Mathematics [NCTM], 1989). The NCTM plays an influential role in mathematical education in the USA. In their most recent educational recommendations published in 2000, they have ventured into describing what constitutes good mathematics education (National Council of Teachers of Mathematics [NCTM], 2000). The document outlines ten standards which are considered to be descriptions of what mathematics instruction should enable students to know (the content standards) and to do (the process standards). There are five content standards: “Number and Operations, Algebra, Geometry, Measurement and Data Analysis and Probability” (p. 29). The other five are process standards: “Problem Solving, Reasoning and Proof, Communication, Connections, and Representation” (p. 29). The process standards are described as ways of acquiring and using content knowledge. It is noticeable how problem solving is considered here as equal with learning to work with representations, communication, and so on, instead of being the goals of these processes (Stacey, 2005).

Similar to the UK and the USA, in Australia, one of the goals mentioned in the 1990 National Statement of Mathematics for Australian Schools was that “students should develop their capacity to use mathematics in solving problems individually and collaboratively” (Australian Education Council, 1990, p. 20). In 1994 the national mathematics curriculum was rewritten in terms of outcomes and was presented as a National Profile (Australian Education Council, 1994). This National Profile “organized the goals of mathematics education into six major strands: five “content” strands plus a “process” strand entitled working mathematically” (Stacey, 2005, p. 343). “Using problem solving strategies” was one of the sub-strands of the process strand. The recent Australian curriculum for mathematics published in 2016 mentions developing ability to pose and solve problems within its second aim. The mathematics curriculum is organised around the interaction of three content strands (“Number and Algebra”, “Measurement and Geometry”, and “Statistics and Probability”) and four proficiency strands (“Understanding”, “Fluency”, “Problem Solving”, and “Reasoning”). So here problem solving is regarded as a proficiency strand, and also considered as one of the “key ideas” (Australian Curriculum Assessment and Reporting Authority [ACARA], 2016). The actions which students can engage in while learning and using the contents relevant to problem solving are described as:

Students develop the ability to make choices, interpret, formulate, model and investigate problem situations, and communicate solutions effectively. Students formulate and solve problems when they use mathematics to represent unfamiliar or meaningful situations, when they design investigations and plan their approaches, when they apply their existing strategies to seek solutions, and when they verify that their answers are

reasonable. (Australian Curriculum Assessment and Reporting Authority [ACARA], 2016, p. 5)

In addition to the central curriculum, problem solving has also been given significant importance in the State mathematics curriculums of Australia (Stacey, 2005).

However, researchers have found that “problem solving activities were almost entirely absent” (p. 487) from certain Australian classrooms (Clarke, Goos, & Morony, 2007). Instead, some research (e.g., Stacey, 2003) has characterised classroom practices as “shallow teaching, where students are asked to follow procedures without reasons” (p. 119).

2.6 Teaching of mathematical problem solving

As discussed in the previous section, mathematical problem solving has been given ample importance in the education systems of different countries over the last few years. How it is supposed to be taught to the students is another crucial aspect. Leading approaches to teaching problem solving include problem based learning (PBL), that is, using problem solving to teach mathematics, osmosis, structured promotion, and specific teaching. If we consider how problem solving is taught in Australia, Clarke et al. (2007) suggest that the three categories mentioned by Schroeder and Lester (1989) - “teaching for problem solving, teaching about problem solving, and teaching through problem solving” (as cited in Stacey, 2005, p. 345) - adequately summarise the three main approaches taken up by the curriculum developers in Australia. Clarke et al. (2007) further suggest that “teaching for problem solving” is reflected in the expansion of the typical curriculum to accommodate the

use of traditional mathematical content to more complex or unfamiliar situations. In a similar fashion, “teaching about problem solving” and “teaching through problem solving” is reflected in specifically including “problem solving heuristics” in the teaching learning processes and introducing novel instruction methods like “problem-based learning (PBL)”. Stacey (2005) studied how problem solving is positioned in different education systems. While focusing on “how the official curriculum materials place problem solving within teaching and assessment activities” (p. 346), she found that the dominant recommendation is to “teach through problem solving”. This essentially means that although mathematical problem solving is not presented as a separate subject area, it is to be taught as a skill through other content areas.

This suggests that research on mathematical problem solving should also focus on the processes rather than the products or results. Based on his research, Mayer (1982) suggested that instead of focusing on the answers, improvements in thinking are achieved by emphasising the processes through which the answer is produced. Focusing on the processes would make it easier to identify which processes students are using to solve a mathematical problem, if they are having difficulty with some of the processes, and how to remedy those problems. This means developing diagnostic techniques so teachers could easily collect information on the status of students’ mathematical problem solving skills. This has the potential of helping teachers better understand the processes that their students are struggling with and helping them to intervene in their learning as necessary. Carpenter, Fennema, Peterson, Chiang, and Loef (1989) found that when teachers were familiar with the cognitive processes students used for mathematical problem solving, their teaching of this area changed significantly where they became more attentive to the methods students use in

problem solving, and were more sensitive to the difficulties students had in understanding the problems.

2.7 Specific area within Mathematical Problem Solving: Arithmetic Word Problem Solving

Mathematical problem solving is a domain-specific area of problem solving that has been researched extensively over the last few decades. There are many areas within mathematical problem solving that different studies have focused on. In order to make this research both in-depth and manageable, it was decided that a specific area within mathematical problem solving needs to be chosen.

Arithmetic word problem solving is an area within mathematical problem solving which plays an important role in the mathematics program at elementary school. In contrast to problems that are presented with mathematical notations, arithmetic word problems are presented through texts where a situation is described, and a question is posed. To answer, students need to understand the situation, identify the numbers and operation(s) necessary for solving the problem, and conduct the operation to find the answer. Below is an example of an arithmetic word problem:

Jukka, Heikki and Pekka went fishing for a day. They had in total 15 sausages and 32 crackers with them. Jukka and Heikki ate, in total, eight crackers for lunch, but big Pekka needed six crackers. The rest of the crackers, they divided equally for supper together. How many crackers did each boy get in the evening? (Pongsakdi et al., 2020, p. 42)

These types of problem solving are often used to help children learn how to apply the formal mathematical knowledge and skills they acquire in the classroom in real-life situations and to exhibit their mastery in previously learned mathematical procedures and concepts. They are considered to be helpful in consolidating students' general problem solving capability, in designing more enjoyable and engaging mathematics instructions, and are utilised in the initial stages of learning a particular concept or skill, for instance, basic arithmetic operations (Thevenot & Barrouillet, 2015; Verschaffel & De Corte, 1997). Although arithmetic word problem solving plays a pervasive role in understanding mathematics, this is an area where students experience the greatest difficulty (Thevenot & Barrouillet, 2015). Earlier research has suggested that such difficulties with word problems have resulted in poor performances in international assessments like TIMMS in some countries (Cheng, 2015; Kawamata, 2007 as cited in Kawasaki, Morita, Umezawa, & Koike, 2013).

Arithmetic word problems might look simple enough, but there are certain complexities integrated within them which make them an interesting area of research. Even when two arithmetic problems represent the same numerical calculation, they might be of varying difficulty because of their semantic characteristics, that is, the way they have been presented (Riley, Greeno, & Heller, 1984; Thevenot & Barrouillet, 2015). Children find it easier to solve problems when the quantities are presented in a straightforward manner compared to when they need to transform the presented situation and then do calculations (Wong & Ho, 2017). For example, in a question like "John has 2 marbles, Tom has 5 marbles, how many marbles do they have together?" it is less difficult for them to find the answer compared to a question like "John has 2 marbles, Tom has 3 marbles more than him, how many marbles does Tom have?" –

though both of these items require a simple addition operation. Such difficulties are attributed to the semantic characteristics of the problems. In addition to the semantic characteristics, there are other factors that play a crucial role in making a word problem easy or difficult. Kintsch and Greeno (1985) studied how children use different type of schemas, first to identify how to solve the problem and then apply another schema to solve it. They suggest that phrases like “more than”, “less than” in the problem statement would trigger a schema which includes “a large set” and “a small set”. Then a difference strategy (that assigns roles based on the more-than/less-than schemata for comparing the two numbers) would be triggered to allow comparisons. Lewis and Mayer (1987) and Fuson, Carroll, and Landis (1996) researched consistent and inconsistent problems and how children apply schemas while solving them. Consistent problems are those where the sentence structure of the word problem is consistent with the operation, for example, “Brenda has 5 apples, Cathy has 6 apples more than Brenda, how many apples does Cathy have?” – here an addition schema is triggered by the “more than” phrase and it is appropriate for this problem. On the other hand, inconsistent problems are those where the sentence might misguide the solver to trigger an incorrect schema, and to be successful the solver has to rearrange the problem, for example, “Brenda has 5 apples, she has 6 apples less than Cathy, how many apples does Cathy have?” – in this problem, initially a subtraction schema is triggered by the “less than” phrase but the problem cannot be solved correctly through this. The solver has to rearrange the problem before solving it. The aforementioned researchers suggest that there is a progression of schema use where at the initial levels the solvers depend on the schemas triggered by

the phrases in problem sentence and gradually get skilled in transforming the inconsistent problems before solving them.

Bassok, Chase, and Martin (1998) and Coquin-Viennot and Moreau (2003) studied how problem solvers apply a situation model in solving problems. Sometimes such problems are phrased in a way that schemas are not readily available for them and the solvers' real-world knowledge plays a crucial role. For example, in the problem 'Mary is giving fruits to guests in an event. She will give 3 apples, 2 oranges, and 1 pear to each of the 20 guests. How many fruits would she need in total?' – children would most likely be solving this problem through a distributed strategy i.e., $(3 \times 20) + (2 \times 20) + (1 \times 20)$. But if the problem is rephrased as 'Mary is preparing fruit baskets for 20 visitors to an event. She puts 3 apples, 2 oranges and 1 pear in each basket. How many fruits would she need in total?', the solution changes to a factorised strategy, that is, $20 \times (3 + 2 + 1)$. The inclusion of 'fruit basket' in the problem sentence triggers a real-life understanding of the situation and changes their solving strategy.

Thevenot, Barrouillet, and Fayol (2004) and Thevenot, Devidal, Barrouillet, and Fayol (2007) explored how posing the question first in the problem sentence facilitates the problem solution. They found that it plays a facilitatory role in encoding and integration of the information presented in the problem. They suggested the findings were not compatible with schema theories as weaker children had more benefit from this practice than those who were usually good at problem solving. If posing the question in the beginning was helping to solve the problem by triggering

the appropriate schema, better problem solvers would find it more useful, as they presumably have a broader range of schemas.

Research on arithmetic problem solving in the past have heavily depended on schema theories for identifying the difficulties students face. But as some of the research discussed above has shown, that is not always the case and there are other sources of difficulty that need to be considered. Thevenot and Barrouillet (2015) suggest that arithmetic word problem solving is a complex activity that requires a series of cognitive processes. There has not been any empirical study that specifically explores the cognitive processes/stages involved in solving arithmetic word problems. Given the important role arithmetic word problem solving plays in students' understanding of mathematics, it was decided that this study is going to focus on this area.

2.8 Developing a theoretical framework for arithmetic word problem solving

This research aims to develop a theory that helps in determining specific difficulties students face while solving Arithmetic Word Problems. In order to understand the stages of Arithmetic Word Problem Solving (AWPS), existing literature was reviewed. The review revolved around theories and frameworks that discuss Mathematical Problem Solving and looked at previous researchers' depictions of how Arithmetic Problems are solved. This section will give a description of how the theories presented by different researchers can be considered for understanding different steps of arithmetic problem solving. It is expected that this discussion will help in constructing a theoretical framework suitable for arithmetic word problem solving.

2.8.1 Theories of Mathematical Problem Solving

In discussing the different theories of problem solving and mathematical problem solving, steps from different models that are similar are discussed together here for ease of comparison.

One of the seminal works in the field of problem solving is George Polya's 1973 book titled *How to solve it*. In this book the author mentions a problem solving method that has four phases. The first phase in this model is "understand the problem" (Polya, 1973). Partially this means clearly seeing what is required, but the problem solver must also desire its solution. The verbal statement of the problem must be understood and the problem solver should be able to identify principal parts of the problem (i.e., the unknown, the data, the condition). Schoenfeld (1985) depicts a series of 'episodes' of mathematical problem solving. Though many of his works related to mathematical problem solving were inspired by Polya, the framework he suggests is more elaborate than that of Polya (1973). It has six episodes, the first of which is "reading". It includes the problem solver reading the problem, absorbing the conditions, contemplating the problem, and perhaps rereading or verbalisation of parts of the problem statement. The second episode in Schoenfeld's (1985) model can also be regarded as part of understanding the problem. It is called "analysis" and occurs if there is no obvious way of progressing once the problem has been read. Analysis is an attempt to understand the problem fully by selecting an appropriate perspective, reformulating or simplifying the problem according to revised perspective and introducing principles or mechanisms that seem appropriate. This often leads to plan development.

The third episode of Schoenfeld's (1985) model, "exploration" can also be considered as a part of understanding the problem. It is different from analysis in the sense that it is more distanced from the original problem and includes a broader visit of "the problem space" to identify suitable information that can help with the solution planning. It might even lead to returning to analysing the problem, using newly gathered information to better understand the problem. Schoenfeld (1985) suggests that a problem solver might find a range of "problem-solving heuristics – the examination of related problems, the use of analogies and so forth" (p. 298) in the exploration episode.

Synthesising earlier research on mathematical problem solving, Artz and Armour-Thomas (1992) proposed a model that has eight steps. They designed it with group problem solving in mind. The model was inspired by Polya's (1973) and Schoenfeld's (1985) work and the steps are quite similar. They added another step to the initial encounter with the problem which they termed "understanding the problem". They suggest that while solving problems, students make frequent comments about the conditions of the problem, which prompted them to include this as the second episode in their model. No further description of this episode is given, hence it is not clear how it is different from Polya's (1973) "understand the problem" phase.

Mayer and Wittrock (2006) suggested four component cognitive processes of problem solving. They term the initial stage of problem solving as "representing" where the problem solver converts the external problem into an internal mental representation. For word problems this would mean developing an internal situation

model that reflects the perimeters of the problem. They suggest that according to classic theories of problem solving, this stage involves constructing a problem space that depicts the initial state of the problem, the goal state, and all legal states in-between.

Wu (2003) developed and validated a theory of mathematical problem solving where she identified four dimensions of mathematical problem solving. The first dimension in her framework is “reading/extracting all information from the question”. She emphasises the importance of understanding the problem before proceeding to solve it. This dimension “goes beyond reading comprehension of the words. It includes understanding the problem situation” (Wu & Adams, 2006, p. 97). That is, understanding the explicit and implicit constraints and assumptions in the problem is part of the process. Within this dimension she highlights a number of aspects of language that influence problem solving, such as general verbal proficiency and technical language proficiency. The fourth dimension in Wu’s framework, titled “real-life and common-sense approach to solving problems” can also be considered to be related to understanding the problem. This dimension discusses how problem solvers sometimes do not comprehend the problem properly because of their inability to bring real world knowledge to the context of solving mathematical problems and emphasises the importance of such knowledge and context. For example, if someone is asked to solve a problem that deals with number of buses required to transport students, he/she needs to have the sense that buses cannot be fractions and consequently the answer would be a whole number.

The second dimension in Wu's model "mathematics concepts, mathematization and reasoning" deals with how a problem situation is converted to an appropriate mathematical formulation. Part of this dimension is similar to Mayer and Wittrock's (2006) "representation" state. Wu (2003) mentions two types of mathematisation that take place in mathematical problem solving: horizontal mathematisation and vertical mathematisation, where horizontal mathematisation is "the process of turning a perceived problem situation into mathematical symbols" (p. 115) (which reflects "representation"), and vertical mathematization is "manipulating the symbols by applying mathematical tools and refining the mathematical models" (p. 115). This dimension seems to cover the mapping of conceptual relations to quantitative representations and at the same time utilising reasoning skills for reorganising mathematics within the mathematical world. Consequently, this dimension does not reflect problem understanding or solution planning stages exclusively. It might be more suitable to think about this dimension as an in-between phase between those two stages mentioned in other models.

"Devising a plan" is depicted as the second phase in Polya's (1973) model. This means knowing, or at least having an outline of the calculations, computations, or constructions that ought to be performed to obtain the unknown (i.e., the solution). This plan might develop gradually, might come after few failed trials, or might come as a sudden realisation, as in "in a flash". Successfully devising a plan requires some existing knowledge of the subject area and concentration on the purpose of problem solving, among other things. Schoenfeld (1985) depicts "planning" as the fourth episode of his problem solving model and emphasises the plan being well-structured,

relevant, and appropriate. Though it is a separate episode, it is not discussed independently to a great extent.

Mayer and Wittrock (2006) suggest that planning and monitoring goes on simultaneously while problem solving; hence the second component process in their model is “planning/monitoring”. Whereas planning entails devising a method for solving the problem, monitoring evaluates the appropriateness and effectiveness of the solution method.

Polya’s (1973) third phase of problem solving is “carrying out the plan”. In this phase the problem solver must check if the details of the problem fit into the solution outlined earlier and then carry out the steps. It is important that the problem solver conceives the plan, rather than getting it from an authoritative figure (like a teacher). A similar notion of operationalisation of a plan is depicted as “implementation” episode in Schoenfeld’s (1985) model. He argues that the implementation of the plan must be orderly/structured. In addition, he suggests monitoring/assessment of the processes by the problem solver. Feedback from this monitoring/assessment is then used for revising the planning and assessment. While this might seem to be similar to Polya’s (1973) final phase, it is more related to the planning state and gives a sense of the metacognitive processes while solving a problem. In this sense Schoenfeld’s (1985) model is not linear and solvers might go back to a previous stage to revise their actions.

“Executing” is presented as the third component in Mayer and Wittrock’s (2006) model and, as in the models discussed earlier, it means a problem solver actually carrying out the planned operations, for example, a problem solver making

necessary calculations to solve an arithmetic word problem. Dimension four in Wu's (2003) model, "computation/carefulness" cover similar areas. The conceptual understanding is covered in the third dimension of her model whereas the fourth dimension is concerned with computational accuracy. This dimension also highlights carelessness, which is more prominent in computation compared to conceptual understanding. Though carelessness might be more of a personality trait than math proficiency, it affects students' ability to obtain the correct answer.

The final phase in Polya's (1973) model is "looking back". He suggests this to be an important and instructive part of problem solving as it gives the solvers an opportunity to consolidate their knowledge and develop their ability to solve problems. This review of the solution can also provide opportunities to improve the solution. A similar idea is given by Schoenfeld (1985) for the final episode of his problem solving model. He calls it "verification" and suggests the problem solver ought to review the solution, test it if possible and consider assessing the solution, an evaluation of the process or an assessment of confidence in the result.

In the model proposed by Mayer and Wittrock (2006), the final component is termed as "self-regulating" and focuses on instigating, modifying, or sustaining the cognitive activities oriented towards goal attainment.

Working on mathematical problem solving in groups, Artz and Armour-Thomas (1992) included "observation and listening" as their final episode. They reasoned that this helps in capturing students' activity of watching and listening to one another.

2.8.2 Theories of Arithmetic Problem Solving

The theories discussed so far are theories of mathematical problem solving. While they have efficiently helped researchers understand different subdomains of mathematical problem solving over the last few years, it might be worth exploring other theories/frameworks that specifically deal with the sub-domain/area this research intends to focus upon, that is, arithmetic problem solving (arithmetic word problem solving). Since this is a content area within mathematical problem solving, it is understandable that theories of problem solving in this sub-domain will have commonalities with theories of mathematical problem solving. Two theories will be discussed here, one developed by Riley et al. (1984) and the other by Thevenot and Barrouillet (2015). Riley et al. (1984) depicted three broad stages of successful arithmetic word problem solving. These stages are based on analysis of computer simulation models. Thevenot and Barrouillet (2015) identified a series of cognitive processes that impact individuals' arithmetic word problem performance.

Closely resembling the theories discussed earlier, the first stage in Riley et al.'s (1984) model is "comprehension". In this stage, problem schemata are used to develop a problem representation from the problem text. Fitting the components of the problem into a coherent structure helps in understanding the word problem. In essence, in my view, this actually has two activities, conceptually understanding the problem and quantification of it. The first stage in Thevenot and Barrouillet's (2015) theory is also "comprehension of a text".

The second stage in Riley et al.'s (1984) model involves "mapping from conceptual relations to quantitative procedures". In this stage the schematised

problem representation is converted to schematised action representation. In my understanding this means the quantified/ numerical conceptual relations are mapped to quantitative procedures in this stage.

Thevenot and Barrouillet (2015) describe two alternative second steps, “the instantiation of abstract representation based on current information” (when enough information is available for constructing a representation) or “the construction of ad-hoc mental models”. Their next (third) step is “reasoning activities to select appropriate arithmetic operations”. Based on the abstract representation or the ad-hoc mental model, problem solvers identify the arithmetic operation most suitable for solving the problem.

“Execution of procedures” is listed as the third and final stage in the Riley et al.’s (1984) framework. In this stage, schematised action representation is converted to problem solution. Thevenot and Barrouillet (2015) term their fourth and final process as “selection, running, and control of algorithmic procedures” which often requires arithmetic fact retrieval from long-term memory and storing intermediary results temporarily. This also seems to be quite similar to Riley et al.’s (1984) final step. It is notable how none of these frameworks/theories has an explicit stage for communication of the problem solution.

2.8.3 Strategies in solving arithmetic word problems

In addition to the steps discussed above, Riley et al. (1984) mention the role of strategy in solving arithmetic word problems. The computer models they used in their research had “strategic knowledge” in addition to problem schemata and action schemata. They suggest such knowledge is represented by production rules that allow

top-down planning and implementation. Thevenot and Barrouillet (2015) also mention that the activities involved in solving a problem require “selection and control of procedures”. This suggests the strategies (sometimes also called procedures) of solving arithmetic word problems might play an important role in understanding and measuring this.

Bisanz and LeFevre (1990) defined strategy as “a procedure that is invoked in a flexible, goal-oriented manner and that influences the selection and implementation of subsequent procedures” (p. 236). The crucial difference between procedures and strategies seems to be the ability of strategies to help in deciding ensuing procedures. This can sometimes mean coming up with new procedures that are loosely based on existing procedures known to the problem solver. Strategies can be deliberate and conscious or unintentional and unconscious (Ashcraft, 1990). Selection of strategy is usually based on its cost value analysis, the cost being the mental resources invested and the value being the accurate results. Siegler (1999) argues that children select problem solving strategies that yield fast and accurate result on the particular problem they are solving, even though they might be well-versed in other strategies.

Lemaire (2010) developed a conceptual framework that depicts four strategy dimensions: strategy repertoire, strategy selection, execution, and distribution. Strategy repertoire is the set of strategies that people use to accomplish a given task. Strategy selection is the ability of picking a strategy from multiple strategies while solving each problem. Strategy execution refers to the implementation of a strategy and is characterised by speed and precision. Strategy distribution is the proportions of

which available strategies an individual is using.

2.8.4 Error analysis in written problem solving

A considerable amount of research has been conducted on error analysis of written mathematical tasks, which is similar to mathematical word problems. The sources of error mentioned by different researchers might provide valuable insight on processes that go on while solving arithmetic word problems.

Newman (1977 as cited in Clements, 1980) presented a number of hurdles/difficulties that a problem solver has to overcome to successfully solve a written problem. She described these as a hierarchy of error causes where if a problem solver is unable to pass a hurdle, he cannot progress to the next level and cannot acquire the problem solution. The steps in her hierarchy of error are: “Reading, Comprehension, Transformation, Process Skills, and Encoding”. She suggested that when solving a one-step mathematical written problem a person has to “read the problem, then comprehend what he has read, then carry out the transformation from the words to the selection of an appropriate mathematical ‘model’, then apply the necessary process skills, then encode the answer” (Clements, 1980, p. 4).

Error analysis studies following this hierarchy collected data through a method that requires interviewing the pupils. While Newman (1977 as cited in Clements, 1980) and Clements (1980) in their respective studies used a method where the students were not helped over problems – hence they had to stop the interview where the students first faced problems – in Casey’s study (1980 as cited in Clements,

1980) students were helped over problems to identify which of the components were troubling them in solving the mathematics.

2.9 A theoretical framework for assessing Arithmetic Word Problem

Solving

The discussion above provides a brief overview of how problem solving, mathematical problem solving, arithmetic problem solving, and error analysis have been conceptualized by different authors. Based on the review of these different theories, a framework for arithmetic word problem solving containing five steps was hypothesised.

Table 2.1
Connection between the theories discussed and the hypothesised AWPS framework

Phase/episode/concept	Source	Step in the hypothesised AWPS framework
Understand the problem	Polya (1973)	Comprehension of the problem
Reading	Schoenfeld (1985)	
Analysis		
Exploration		
Understanding the problem	Artz and Armour-Thomas (1992)	
Representing	Mayer and Wittrock (2006)	
Reading/extracting all information from the question	Wu (2003)	
Mathematics concepts, mathematization and reasoning*		
Real-life and common sense approach to solving problems*		
Reading the problem	Newman (1977 as cited in Clements, 1980)	
Comprehension of the problem		
Comprehension	Riley et al. (1984)	

Comprehension of a text	Thevenot and Barrouillet (2015)	
Devising a plan	Polya (1973)	
Planning	Schoenfeld (1985)	
Planning/monitoring	Mayer and Wittrock (2006)	
Mathematics concepts, mathematization and reasoning*	Wu (2003)	
Mapping from conceptual relations to quantitative procedures	Riley et al. (1984)	Mapping conceptual relations to a numerical representation
The instantiation of abstract representation based on current information or the construction of ad-hoc mental models	Thevenot and Barrouillet (2015)	
Reasoning activities to select appropriate arithmetic operations*		
Transformation of the problem	Newman (1977 as cited in Clements, 1980)	
Strategic knowledge	Riley et al. (1984)	
Reasoning activities to select appropriate arithmetic operations*	Thevenot and Barrouillet (2015)	Selecting an appropriate strategy for solution
Selection and control of procedures		
Strategy repertoire, strategy selection, execution, and distribution	Lemaire (2010)	
Selection, running and control of algorithmic procedures*	Thevenot and Barrouillet (2015)	
Carrying out the plan	Polya (1973)	
Implementation	Schoenfeld (1985)	
Executing	Mayer and Wittrock (2006)	
Computation/carefulness	Wu (2003)	Performing calculations
Execution of procedures	Riley et al. (1984)	
Process skills	Newman (1977 as cited in Clements, 1980)	
Selection, running and control of algorithmic procedures*	Thevenot and Barrouillet (2015)	
Looking back	Polya (1973)	
Monitoring/assessment of the processes	Schoenfeld (1985)	Verifying the solution and communicating the results
Verification		
Self-regulating	Mayer and Wittrock (2006)	

Real-life and common sense approach to solving problems*	Wu (2003)
Encoding	Newman (1977 as cited in Clements, 1980)

*This concept fed into more than one step of the hypothesised framework

Table 2.1 shows how the phases/steps/concepts of the different existing theories are connected with the steps of the hypothesised framework. Similar concepts presented by different authors were grouped together and consolidated into a single step in the hypothesised framework. In some cases, multiple steps from one theory were grouped together as they represent the same concept in the hypothesised theory. For example, three episodes from Schoenfeld's (1985) theory of mathematical problem solving – *reading, analysis, and exploration* – were considered under the *comprehension of the problem* step in the proposed framework. In addition, some concepts from the existing theories were considered to be feeding in to more than one step of the hypothesised framework. Wu's (2003) theory discusses *mathematics concepts, mathematization and reasoning* as the second dimension of mathematical problem solving. This dimension seemed to include ideas that relate to *comprehension of the problem* as well as *mapping conceptual relations to a numerical representation* step of the hypothesised framework.

The following is an initial hypothesis of the cognitive steps in arithmetic word problem solving. Naturally, it is expected that this theoretical framework is going to be changed based on the analyses of quantitative and qualitative data. The hypothesised steps are:

a) *Comprehension of the problem:* this would include reading the problem, analysing and exploring it to have a clear sense of what the problem is about. Successful completion of this stage would give the problem solver a clear idea about the different elements in the problem and the situation described in it.

b) *Mapping conceptual relations to a numerical representation:* this step would involve constructing a quantitative/numerical mental model of the concepts presented in the problem. This can be regarded as mathematisation of the problem and real-life and common-sense approaches would play a role in this mapping. At this stage the problem solver would have a quantitative/numerical understanding of the initial and the goal state.

c) *Selecting an appropriate strategy for solution:* this step would include the problem solver exploring his/her strategy repertoire (mental collection of procedures) to identify an appropriate and effective method for solving the problem. If none of the existing strategies are readily suitable for the problem, they might retrofit an existing strategy with concepts they have obtained from other strategies and evaluate its suitability in solving the problem they are trying to solve.

d) *Performing the calculations on the stored numbers:* in this step the problem solver would execute the plan he/she had made by using the procedure selected. This would mean using the selected strategy/procedure on the quantitative representation of the initial stage to reach the goal stage. Carefulness in carrying out the computation would be an important element of

this stage. There may also be low stress methods they try using, make a guess and check.

e) *Verifying the solution and communicating the results:* this stage would involve reviewing the solution, evaluating its appropriateness and checking for errors. This would provide cognitive feedback to the problem solver regarding the success/failure of the journey through previous processes. Furthermore, it would include communicating the results in an appropriate format.

Though these are processes which take place one after another, (a) sometimes they may appear so quickly that it appears to be simultaneous, and (b) the completion of all the steps might not always be unilateral (i.e., start and go forward without going back). Sometimes a problem solver might get back to an earlier step after developing some new understandings from a later state.

This abovementioned discussion so far essentially summarises how a hypothesised framework on the cognitive processes of students' arithmetic word problem solving activity has been developed. This framework would help in finding ways of assessing the construct of interest in this study, that is, arithmetic word problem solving. After delineating the construct the next step is to identify the behaviours, products or actions which will help in gathering evidence to draw inferences about it (Wilson, 2002). In order to make the present study more manageable, only problems which are less likely to be iterative were considered. Furthermore, only arithmetic word problems that require addition and subtraction operations were taken into account and other operation types were excluded. The

next chapter of this thesis discusses how items were developed, data were collected and analysed to determine the validity and reliability of the proposed framework. But before that, the next sections of this chapter outline how existing literature was reviewed to determine the suitable approaches to validity and reliability for this research.

2.10 Approaches to validity and reliability

The following sections discuss the theories of validity and reliability that helped in forming the understanding of the concepts used in this study. Definitions of validity and validation are discussed first. Then an overview of how these concepts evolved over the years, and how best to perceive them for the current study are given. Finally, a brief discussion of reliability concepts is presented at the end.

2.10.1 Validity

Validity and validation are important concepts in developing assessments. Validity of a test is sometimes defined as “the test’s ability to measure what it is supposed to measure or portrays to measure” (Azim, 2018, p. 225). However, as further discussions here will show, this definition is overly simplistic and does not take into account more recent developments in validity theory. Sireci (2009) suggests that such simplistic definitions of validity were pragmatic and focused on correlation of test score with some known criterion. A discussion of how validity theories evolved over the years (presented later in this text) explains this better. According to the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014), “Validity refers to the degree to which

evidence and theory support the interpretations of test scores for proposed uses of tests” (p. 11). It is noticeable how in this definition the focus is on the *degree* of alignment and on the interpretations of the scores for specific uses, rather than on the test itself. Perhaps the most widely used definition of validity characterises it as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment” [italics in original] (Messick, 1989, p. 13). It is noticeable that the definition from the 2014 standards (presented above) is also a reflection of Messick’s definition.

Validity is not an innate trait of a test, rather it is a way of validating the inferences derived from the test scores (Sireci, 2009), that is, whether there is evidence for the validity of the instrument usage (Wilson, 2005).

Validation on the other hand is the process of collecting validity evidence. Cronbach (1971) described validation as the process used by a test developer or test user to collect information to support the planned inferences of the test scores. Similarly, Crocker and Algina (2008) defined validation as a process of gathering empirical evidence for supporting test score use for a given purpose.

A review of literature suggests there has been a distinct shift in how validity is conceptualised over the decades. Initially, validity was classified into categorically different types (e.g., criterion-related validity, content validity, etc.), which changed to a multifaceted but unitary concept of validity over the years. For example, Ebel (1956) presented four types of validity – “content validity, concurrent validity, predictive validity, and construct validity” (p. 269). However, recent literature discusses validity

as a unitary concept with different types of evidence supporting it (Reeves & Marbach-Ad, 2016).

In the early days of validity theory, criterion-related validity emerged with the proliferation of statistical methods, such as correlation and factor analysis, that helped in determining whether tests were measuring what they were supposed to measure (Sireci, 2009). Performance in a test was compared with some existing criteria that was believed to be relevant to the test's area of focus, and validity was determined based on the degree of empirical relationship (determined through correlation or regression) (Messick, 1989). This was regarded as the gold standard for assessing validity (Kane, 2006). There were two types of criterion-related validity in use, concurrent validity, where relationship of test score with some current criterion performance was compared, and predictive validity, where the test score's relation with some future behaviour was determined (Lissitz & Samuelsens, 2007). This model worked well when there was a criterion that can be easily compared. However, it was not possible to use this model where an acceptable criterion was not available (Kane, 2006).

Content validity appeared as the dissatisfaction with correlation-based criterion validity increased and focused on both the test content and the criterion against which they were correlated (Sireci, 2009). The content model used a sample of performance in an area of interest, a sample that is representative of the domain (Kane, 2006). Professional judgment about the relevance and representativeness of the test content for an area of interest (e.g., a behavioural domain) was used to determine content validity (Messick, 1989). Another approach was to use a test plan

or blueprint to ensure selection of representative sample of tasks/performances from the domain of interest. This model has been used abundantly for measuring academic achievement (Kane, 2006).

During early to mid-1900s, educators/researchers were primarily working with a criterion-based validity model with some concentration on content-based validity (Zumbo, 2009). These models worked for workplace assessments (e.g., job placement) or achievement tests but they were not suitable for developing tests that assessed psychological attributes for which there were neither an existing criterion to compare to, nor a performance domain from which to sample tasks/performances. Hence, there was need for a model that can validate measures of theoretical attributes (Kane, 2006). That is where the concept of construct validity came in. “Construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not ‘operationally defined’” (Cronbach & Meehl, 1955, p. 251 emphasis in original). In addition to taking into account the qualities (constructs) that are assessed, this model considers the nomological network of how these qualities are related to one another (Lissitz & Samuelsen, 2007; Sireci, 2009). The meaning of construct validity has evolved over the years and in its dominant current use it is considered as “the degree to which inferences can be made legitimately from the observed scores to the theoretical constructs about which these observations are supposed to contain information” (Zumbo, 2009, p. 68). According to (Kane, 2006), construct validity was framed based on the “hypothetico-deductive (HD) model of scientific theories” by Cronbach and Meehl (1955). In this model, “a theory consists of a network of relationships linking theoretical constructs to each other and to

observable attributes. Each theoretical construct is implicitly defined by its role in the theory” (Kane, 2006, p. 20).

This is especially important for this study as, according to Kane (2006), in the construct validity approach, when predictions derived from theory are not reflected in observations, the measurement, the theory or some ancillary assumption of the process is thought to be at fault. But if the predictions are confirmed in the observations, both the theory and the interpretations of scores are considered supportable (i.e., validated). Since this research is about validating a theoretical framework for Arithmetic Word Problem Solving (AWPS), using a construct validity model is suitable for the study. This concept of validating a theory by determining validity of a test that is developed based on the theory is also discussed by Crocker and Algina (2008), Sireci (2009), Zumbo (2009), and Hubley and Zumbo (2011).

Loevinger (1957) argued that “construct validity is the whole of validity from a scientific point of view” (p. 636) as she suggested “predictive, concurrent, and content validities are all essentially ad hoc” (p. 636). Experts took this broad concept of construct validity to be encompassing all evidence for validity (content and criterion evidence included), reliability, and the wide range of theory testing methods, and Messick (1989) utilised the construct model as the basis for a unifying framework of validity (Kane, 2006). This idea became prominent as the unitary concept of validity with construct validity regarded as the overarching concept for the unified validity theory where collection of multiple types of evidence is required for construct validation (Crocker & Algina, 2008; Lissitz & Samuelsen, 2007; Sireci, 2009). So, in the

current conception, construct validity is regarded as the whole of validity theory (Zumbo, 2009).

Even though Messick's unitary concept of validity is widely accepted in current literature, some authors (e.g., Hubley & Zumbo, 2011; Lissitz & Samuelson, 2007) convey their dissatisfaction with this unitary approach as they think it is impractical. Others believe that the idea of construct validation has been elevated to such a high level that it has become an out-of-reach goal (Fremer, 2000, as cited in Lissitz & Samuelson, 2007). Sireci (2009) considers an argument-based approach to validity to be a compromise between the overly simplistic view of validity and validity as an unobtainable goal. While an argument-based approach is still based on the unitary concept of validity, it articulates various types of evidence that can be collected to support validity arguments. It is not an all-or-nothing proposition. Based on the evidence collected, the degree of validity can be determined. Experts like Wilson (2005) and Kane (2006) also suggest using an arguments-based approach where different validity evidence is collected to develop a validity argument. The 2014 Standards for Educational and Psychological Testing (American Educational Research Association [AERA] et al., 2014) also articulate an argument-based approach to validity. Wilson (2005), American Educational Research Association [AERA] et al. (2014), Reeves and Marbach-Ad (2016) and others (e.g., Sireci, 2009) discuss a similar set of validity evidence for an argument-based approach. These can be summarised as follows:

- a) Evidence based on instrument content
- b) Evidence based on response processes

- c) Evidence based on internal structure
- d) Evidence based on relation to other variables
- e) Evidence based on the consequences of using an instrument

The following sections briefly discuss each of these evidence types.

a) **Evidence based on instrument content:** Such evidence is similar to traditional content validity evidence (Sireci, 2009) and depicts the match between the instrument and the domain the instrument purports to measure (Reeves & Marbach-Ad, 2016). Sireci (2009) suggests that content validity evidence includes practice (job) analyses, review by subject matter experts, test specification, and so on. Reeves and Marbach-Ad (2016) also highlight test review by subject matter experts and using a table of specification/test blueprint while designing the test as suitable validity evidence of this type. According to Wilson (2005) collecting such evidence would entail (i) defining the construct (preparing the construct map), (ii) describing the set of items that comprise the test (preparing the item design), (iii) deciding on a strategy to code the responses into an outcome space (preparing the outcome space), and (iv) preparing a technically calibrated version of the construct (the Wright map).

b) **Evidence based on response processes:** Such evidence determines whether the performance of examinees is corresponding with the assessment of the construct (Sireci, 2009). In order to collect evidence based on response processes, the test developer needs to analyse individual response processes while they are taking the test or just after taking the test in an exit interview

(Wilson, 2005). American Educational Research Association [AERA] et al. (2014) suggest that it is helpful to interview test takers from various groups of intended test takers about the strategies they followed while answering the items to gain information that enrich the definition of the construct. Sireci (2009) also outlines post-test interviews about response processes, response behaviour observation, and analysis of response time data as suitable evidence about response processes. Such validity evidence is often gathered through think-aloud protocols administered with the respondents (Reeves & Marbach-Ad, 2016).

c) **Evidence based on internal structure:** This type of evidence involves item data to determine the dimensions measured by an assessment (Sireci, 2009). As Wilson (2005) suggests, to gather evidence based on internal structure, there needs to be an intended internal structure of the instrument. If the construct has been built using the previously mentioned steps (under evidence based on instrument content), there will be an expectation of internal structure with the construct being in certain order (high to low, easy to difficult, etc.). He further elaborates that one type of evidence based on the internal structure is gathered through analysis of test data using a measurement model and displayed in a Wright map. The comparison of the theoretically expected levels and the empirical data-based levels demonstrates the support or contradiction that the evidence offers. Other evidence based on internal structure includes demonstrating that the items are consistent with the instrument as a whole. Another source of evidence is analysis of differential item functioning (Wilson, 2005). Presenting a similar concept,

Reeves and Marbach-Ad (2016) discuss that fidelity of test scores to their theorised internal structure can be established through “techniques such as factor analysis, item response theory and Rasch modelling” (p. 4). They suggest that, in determining whether the internal structure of the test reflects the theorised internal structure, model-based predictions are compared with observed data from the test, checking if the model fits the data. These ideas conform with the American Educational Research Association [AERA] et al. (2014) discussion of evidence based on internal structure as well.

d) **Evidence based on relations to other variables:** Traditionally predictive and concurrent criterion relationships have been used to determine test-criterion relationships (American Educational Research Association [AERA] et al., 2014), which was termed as criterion-validity. Evidence based on relations to other variables reflects traditional criterion validity evidence (Sireci, 2009). In cases where there are external variables that should be related to the construct of interest, and there are instruments that are supposed to measure the same or a similar variable, the empirical relation between the test that is being validated and these external variables can be utilised as validity evidence (Wilson, 2005). Collection of this type of validity evidence focuses on determining test scores’ relation with other measures of same or similar construct (termed convergent validity evidence) and measures of distinct and different constructs (termed discriminant validity evidence) (American Educational Research Association [AERA] et al., 2014; Reeves & Marbach-Ad, 2016). Associational methods (e.g., correlation analysis) are often utilised to determine relationship of test scores with the external variables

(Reeves & Marbach-Ad, 2016; Wilson, 2005). Examples of external variables could be scores on other tests, teacher ratings, grades, and so on (Wilson, 2005). It is very important to remember that the quality of such validity evidence (based on external variables) depends on the quality of the external measure used (Reeves & Marbach-Ad, 2016). It ought to be mentioned that, in gathering this type of validity evidence, analysis of a predictive criterion relationship determines strength of correlation between the test in focus and some criterion scores obtained afterwards, whereas concurrent criterion relationships are determined by analysing correlation between the test and an existing measure of the same or similar construct (American Educational Research Association [AERA] et al., 2014).

e) **Evidence based on consequences of testing:** According to Sireci (2009), “evidence based on consequences of testing refers to evaluation of the intended and unintended consequences associated with a testing program” (p. 31). However, American Educational Research Association [AERA] et al. (2014) depicts this type of evidence as consequences of test use that “follow directly from the interpretation of test scores for uses intended by the test developer” (p. 19). That is, they exclude consequences arising from unintended use of the test scores. An example of unintended consequences of testing is the narrowing of curriculum that occurred due to the No Child Left Behind Act testing regime (Reeves & Marbach-Ad, 2016). Other examples of such evidence could be data on adverse impact of the test, effects of the test on other issues such as school dropout or job application, and so on (Sireci, 2009). Surveys, interviews, focus group discussion, and other methods are often used to

gather evidence based on the consequences of testing (Reeves & Marbach-Ad, 2016). It is important to remember that, if a particular instrument has negative consequences, even with reasonable validity evidence from the other sources, its use should be reconsidered (Wilson, 2005).

2.10.2 Developing a coherent validity argument based on different types of evidence

Whether or not a test is valid for a purpose is always going to be a matter of judgment. Having a sound validity argument based on a range of evidence makes it easier for different audiences to come to the same conclusion regarding the validity (Sireci, 2009). The proposition here is for a validity argument that brings together various sources of validity evidence into one coherent account. This argument is used to determine “the degree of validity”, that is, how the collected evidence and theory support the intended interpretation of the test score. In some cases, the validation process might demonstrate a need for redefining the construct or revising the test, which might require further studies. Validation processes are often iterative, as there is always further information to be collected and integrated in the validity argument (American Educational Research Association [AERA] et al., 2014; Wilson, 2005).

2.10.3 Reliability

Reliability is another important characteristic of assessment. Traditionally, it has been viewed as a quality separate from validity but in the current conception it is regarded as an integral part of validity (Wilson, 2005). It is sometimes considered as “a necessary but not sufficient condition for validity” (Miller, Linn, & Gronlund, 2009, p.

108). Comparing validity and reliability, Haertel (2006) suggests that “whereas validity is centrally concerned with the nature of the attributes tests measure, reliability is concerned solely with how the scores resulting from a measurement procedure would be expected to vary across replications of that procedure” (p. 65). Kane (2006) suggests that the broad conception of construct validity encompasses reliability, along with the different types of validities (as discussed above). It is generally accepted that reliability of an instrument is related to the different validity characteristics (American Educational Research Association [AERA] et al., 2014; Wilson, 2005).

In simple terms, reliability indicates the extent to which assessment results are replicable (Wu, Tam, & Jen, 2016). Cronbach (1951) defined reliability as the accuracy or dependability of measurements. Similarly, Miller et al. (2009) suggest reliability refers to how consistent assessment results are from one measurement to another, that is, it is an indication of the consistency of the assessment scores. Reliability is considered to be a fundamental characteristic of a test, not dependent on any external measures for its meaning. However, the level of reliability coefficient is partially a function of the respondent sample for the test (Lissitz & Samuelson, 2007).

Definition and derivations of test reliability forms the core concepts of classical test theory (Adams, 2005; Wu et al., 2016). Nevertheless, reliability can be calculated in both Classical Test Theory (CTT) and Item Response Theory (IRT) approaches. Classical test theory revolves around the concept of true score where any test gives us an observed score that is a function of the true score of the respondent and some measurement error ($X = T + E$). In theory, administering the same test again and again with the same respondents, where the respondents are made to forget about the test

content every time they take the test, and then averaging the observed scores from the multiple administrations, gets us closer to the respondents' true scores. Reliability depends on how much measurement error comes through in the observed score. It can be expressed as "the ratio of the variance of the true scores over the variance of the observed scores across the population of students" (Wu et al., 2016, p. 10). Since observed score is composed of the true score and measurement error, "if the variance of error scores is small compared to the variance of the true score, reliability will be high" (Wu et al., 2016, p. 10). Similarly, if the variance of error score is high in comparison, reliability is going to be low.

Miller et al. (2009) depict the main methods of assessing reliability as: test-retest (as a measure of stability), equivalent-forms (measure of equivalence); test-retest with equivalent-forms (measure of stability and equivalence), split-half (measure of internal consistency), coefficient alpha (measure of internal consistency) and interrater (measure of consistency of ratings). In IRT approaches, where students' raw scores are interpreted through measurement models, traditional approaches of determining reliability have been "to compare how much of the observed variance in respondent locations is attributable to the model as a proportion of the total variance" (Wilson, 2005, p. 145). Wilson (2005) also highlights the different methods that can be taken to determine reliability in different approaches:

- (a) internal consistency (considers the proportion of variance the model has accounted for),
- (b) test-retest (considers consistency of measures over time),
- (c) alternate forms (considers different sets/forms of items), and

(d) consistency between raters (when test is scored by raters).

Calculations of reliability or consistency in the first three of these approaches are based on the measurement error of the scores.

When one set of items is administered only once, a measure of internal consistency is the most suitable way of determining reliability. In the CTT approach (using raw scores), the split-half method has been used by educators to calculate the correlation coefficient between the two halves of the test as a method of determining reliability (Miller et al., 2009). Probably the simplest way of splitting a test is to separate the odd and even number items and compare them through correlation coefficient. However, it is statistically more sound to use an approach where all the possible combinations of split-halves are considered in calculating the final correlation coefficient. Cronbach's alpha coefficient has been shown to be an estimate of the mean reliability of all possible splits (Wu et al., 2016). Using Cronbach's alpha removes error associated with an arbitrary choice of splits (Neumann, Neumann, & Nehm, 2011) and it is "the most widely used coefficient of internal consistency" (Haertel, 2006, p. 74). Another approach of calculating the mean reliability based on all possible splits of a test is the KR-20, which has been demonstrated to be "a special case of the Cronbach's alpha coefficient when the latter is applied to dichotomously scored data" (Wu et al., 2016, p. 80).

While using IRT for analysing test data, raw scores are interpreted using a measurement model. When respondent locations are estimated in such models, there is a degree of uncertainty. This uncertainty (of location) is characterised as the standard error of measurement (SEM) and conveys the accuracy of the estimates. If

the SEM value is too large, it is not possible to make sound interpretations of the scores (Wilson, 2005). The SEM can be used to calculate the Person Separation Reliability which is the “ratio of the true variance of the latent variable to the observed variance in the students’ estimates” (Adams, 2005, p. 165).

IRT approaches can utilise different estimation methods for determining a person’s ability, such as WLE (weighted likelihood estimation) and EAP (expected a posteriori). These estimates are used for determining reliability based on the separation reliability formula. For example, as discussed by Adams (2005), WLE reliability is determined based on the “variance of the person ability estimates calculated in the model and the average of the squared errors of these estimates” (Binder et al., 2019, p. 7). Researchers such as Martin and Lazendic (2018) and Binder et al. (2019) have used WLE to calculate reliability. Another IRT method of determining reliability is using the Expected A Posteriori/Plausible Value (EAP/PV) estimations. In this method, predictive reliability is calculated as 1 minus the ratio of the variance of a participant’s score distribution (compared to the sample variance) and EAP/PV reliability is the mean of all such predictive reliabilities from the sample (Neumann et al., 2011; Zoanetti, Griffin, & Adams, 2006).

Chapter summary

This chapter outlined how reviewing literature has helped in delineating the area of research for this study, how the hypothesised framework for assessment of arithmetic word problem solving was developed and how suitable approaches for determining

validity and reliability of the framework were identified. The next chapter discusses the methodology and methods used in this study.

Chapter 3: Methodology and methods

This chapter discusses the methodology and methods used in this research. Once the framework for assessing Arithmetic Word Problem Solving was hypothesised based on existing literature, the next step was to collect empirical evidence to determine the validity of this framework. As discussed in the previous chapter, an arguments-based approach to validation was adopted in this research where different types of validity evidence are collected to determine the validity of an assessment. Existing literature on areas such as developing an assessment and psychometrics give an indication of the processes that can be followed to collect the validity evidence. It is understood that a field of research dictates what is suitable for that research, and some discipline areas are more aligned with certain research philosophies/paradigms than others because of the nature of the knowledge that they deal with. However, with some initial understanding of what the necessary activities are to collect appropriate evidence for this study, the researcher wanted to determine where this research should be placed within the extant research philosophy/paradigm/methodology conceptions.

The first few sections of this chapter briefly discuss literature related to areas such as ontology, epistemology, and describe how pragmatism was considered to be the suitable paradigm for this research. Next, mixed method research methodology is discussed and why it is appropriate for this research is explained. The ensuing sections discuss the different instruments used for collecting different types of validity evidence, how the instruments were developed, how the participants for this research were selected, the data collection process (how the data collection instruments were

administered), the data management process, and the data analysis process. The final section of the chapter discusses the ethical considerations of this research.

3.1 Research philosophy

With the advancements in the theory and practice of social science research in general and educational research in particular, researchers now have a range of approaches that can be utilised to conduct a study. These research approaches are reflections of different research philosophies and whether, consciously or not, researchers bring certain beliefs and philosophical assumptions to their research (Creswell & Poth, 2018). In most published research, especially journal articles, the philosophical ideas behind the research remain hidden. However, these ideas influence the design and conduct of the research and some careful scrutiny of research reports would often give an indication of which philosophical camp the researchers belonged to. One of the crucial tasks for beginning researchers while conducting a study is to articulate the philosophical position of their research. While it is sometimes troublesome for them to gather an understanding of the philosophy that underlies the design and implementation of their research, it is imperative that they gain the necessary knowledge on this field and decide on their philosophical position before commencing their research journey. The reasons why understanding and articulating the philosophy of research is so important are, first, how the research is conducted is influenced by the philosophical assumptions (Creswell & Poth, 2018; Denscombe, 2010). The perspective adopted regarding the research topic, formulation of problem and research questions, nature of investigation, methods, and so on, are all shaped by the research philosophy. In addition, research philosophy determines what constitutes

worthwhile evidence/data, and what conclusions can and cannot be reached based on the investigation. Second, a researcher's philosophies reflect their previous research training, their discipline orientation, and are reinforced by the scholarly community they work in (Creswell & Creswell, 2018; Creswell & Poth, 2018). Third, there are alternative categories of philosophical assumptions supporting different positions, and researchers need to understand these and place their research properly (Denscombe, 2010). Finally, reviewers of research also make philosophical assumptions about studies they are evaluating (Creswell & Poth, 2018). Explicitly mentioning a research's philosophical position makes it easier for the evaluators to criticise it accordingly. Overall, researchers have to embrace a research design based on the philosophical assumptions they themselves align with. This design, that is, the overall blueprint of the research, inclusive of data collection, should focus on ways most conducive to achieving the aims of the research. It should be 'fit for purpose' (Cohen, Manion, & Morrison, 2018).

Philosophical assumptions of research consist of a set of beliefs that guide actions or inquiries. These are often called worldviews or paradigms (Cohen et al., 2018; Creswell & Clark, 2018; Creswell & Creswell, 2018; Kivunja & Kuyini, 2017). According to Cohen et al. (2018), paradigmatic pluralism and multiple ways of constructing paradigms are often observed in social science research. There are some differences and a lot of overlaps among the definitions and elements of paradigms presented by different authors. According to Lincoln and Guba (1985), a paradigm comprises four elements: ontology, epistemology, axiology, and methodology. In a similar fashion, Creswell and Poth (2018) also identify these four elements as four types of philosophical assumptions required in a research study. However, Scotland

(2012) suggests that a paradigm consists of “ontology, epistemology, methodology and methods” (p. 9). Although there might be minor differences in how different authors have articulated the elements of a research paradigm, the key components remain similar. Generally, it is agreed that within a research paradigm, ontological assumptions lead to epistemological assumptions, which lead to instrumentation and data collection (Cohen et al., 2018; Tuli, 2010). The methodological consequences of selecting a paradigm permeate “the research question/s, participants’ selection, data collection instruments and collection procedures, as well as data analysis” (Kivunja & Kuyini, 2017, p. 36). Hence it is important for a researcher to decide on a research paradigm. However, we should also keep in mind that paradigms are not mutually exclusive or strictly separate, there might be overlaps between different paradigms and they can be combined or used separately (Cohen et al., 2018; Creswell & Clark, 2018).

There are a lot of overlaps in the literature where research philosophy, paradigms, worldviews, ontology, epistemology and so on, have been discussed. The same words have been used by some researchers to name paradigms, while others have used them for describing ontologies and others have used them for epistemologies. This is partly because many authors have named the paradigms based on the epistemology they entail. For example, Creswell and Clark (2018) discuss four worldviews: post-positivism, constructivism, transformative worldview, and pragmatism. Creswell and Creswell (2018) also discuss these four worldviews. Cohen et al. (2018) discuss positivism, post-positivism, naturalistic and interpretive approaches, postmodernist and post-structuralist perspectives. Creswell and Poth (2018) discuss post-positivism, social constructivism, transformative frameworks,

postmodern perspectives and pragmatism. Kivunja and Kuyini (2017) outlines four paradigms: positivist, interpretivist/constructivist, critical/transformational, and pragmatic paradigm. Scotland (2012) mentions three paradigms: scientific, interpretive, and critical. Somewhat different from these authors, Denscombe (2010) discusses paradigms based on research approaches. According to him, there are three research paradigms: quantitative, qualitative and mixed methods research paradigms.

Even with the overlaps and different articulations, there is some coherence in the literature in the sense that they all recommend the same epistemological approach for the same ontological stances. As Creswell and Poth (2018) suggest, any discussion of a framework (that relates the components of philosophy of research) can only be a partial description of the possibilities, however, a discussion of some major frameworks can give a general sense of how this works. Keeping this in mind, the following sections will discuss the concepts of ontology, epistemology, axiology and methodology.

3.1.1 Ontology

The focus of ontology is the nature of reality. Creswell and Poth (2018) define ontology as the researcher's view of reality. Similarly, Creswell and Clark (2018) suggest that ontology is the nature of reality, that is, what is "real" according to the researchers' assumption when they conduct their inquiries. Scotland (2012) also posits that ontological assumptions refer to what constitutes reality, that is, focuses on the question, what is? He argues that researchers should decide on their position about their perception of how things really work and how things really are. While these definitions consider reality in general, Denscombe (2010) narrows his definition

down to social reality. According to him, ontology discusses the nature of social phenomena and the researcher's beliefs about the nature of social reality. Denscombe (2010) posits four types of ontology – Realism, Constructionism, Critical Realism, and Pragmatism. Based on his discussion, here is a brief description of how proponents of these ontologies perceive reality:

Realism - realists conceive the social world to be similar to the natural world, having properties that can be measured, and having stable and fairly consistent structures and relationships. The social world is out there as an objective reality independent of individuals believing or approving it or not.

Constructionism - Constructionists think the social world is created by human mind, reality is constructed through perceptions and interactions. Hence it varies among different cultures and groups. In contrast to the realists, constructionists believe the social world to be comprised of multiple realities, so researchers need to focus on exploring how different groups construct their social realities. There is no ultimate, definitive version of social phenomena.

Critical realism - critical realism is regarded as an alternative to positivism and interpretivism, though it has much affinity with positivism. In essence, it advises social researchers to be cautious about their claims and consider their theories to be more tentative than a complete and final account of the world. The main ontological assumptions of critical realism are – reality is not always observable and independent of an individual's experience or interpretation, its impact is not always predictable, and it is complex so not always revealed by measurements and observations.

Pragmatism - Pragmatism stems from the scepticism about the philosophical debates among positivism, interpretivism and critical realism and regards such debates as unproductive. It investigates the suitability of any method or approach on its suitability of outcomes for the research problem in practice. This approach became popular among the social researchers as they have shown a tendency of picking and choosing from a collection of methods at their disposal, often simultaneously using approaches representing positivism and interpretivism. Consequently, they recognise no single approach is perfect for their purpose as they have focused on getting the best from the available tools. Pragmatism as an approach is more readily associated with practitioners doing applied research.

Pragmatism is driven by practicality, instead of ensuring consistency with a specific philosophical position it pursues useful outcomes. It is possible to regard social reality to be external to individuals, that is, being 'out there' and simultaneously consider it as socially constructed, that is, 'in the mind'. The researcher uses a vision of social reality that might prove to be more useful.

3.1.2 Epistemology

Epistemology focuses on the nature of knowledge, how we gather knowledge about our reality and what it means to know. According to Denscombe (2010), it involves the ways in which humans create knowledge about their social world and includes philosophical debates regarding the efficacy of our methods to gather knowledge of social reality. Kivunja and Kuyini (2017) also consider epistemology to represent how we know something – how truth or reality is known. Creswell and Poth (2018) suggest that epistemology outlines how the inquirer knows reality, what can be regarded as

knowledge and how claims about knowledge are justified. Epistemology asks questions like “what is the relationship between the knower and what is known? How do we know what we know? What counts as knowledge?” (Tuli, 2010, p. 99). According to Denscombe (2010) there are two fundamental positions of epistemology: positivism and interpretivism. Kivunja and Kuyini (2017) outline four categories of epistemological basis of research. According to them, for a researcher relying on “forms of knowledge such as beliefs, faith and intuition” the epistemological basis is “intuitive knowledge” (p. 27); for someone relying on data gathered from leaders in organisations, books, people ‘in the know’ – the epistemological basis is “authoritative knowledge” (p. 27); for a researcher who emphasises that reasoning is “the surest path to knowing the truth”, the epistemological basis is “rationalist epistemology or logical knowledge” (p. 27); and if a researcher considers that “knowledge is best derived from sense experiences”, and noticeable, objective facts, then his epistemological basis is “empirical epistemology” (p. 27).

3.1.3 Axiology

Axiology considers the values and beliefs we hold (Cohen et al., 2018) and outlines the discussions of ethical issues that need to be considered while conducting research. According to Kivunja and Kuyini (2017), it addresses the general question “What is the nature of ethics or ethical behaviour?” (p. 28). Others suggest that axiology outlines the nature of value and how the researcher’s values may influence what is being studied (Weaver, 2018). Creswell and Poth (2018) also describe axiology as the value-stance taken by the researcher and the role of values in research.

Based on the proposition that the axiology of a study is concerned with issues of ethics, Kivunja and Kuyini (2017) refer to four criteria of ethical conduct: teleology (focuses on getting to a meaningful outcome of research that satisfies as many people as possible, that is, the research produces more benefit than harm.); deontology (focuses on understanding that each action taken during the research will have its own consequences); morality (focuses on ensuring that intrinsic moral values be upheld during the research – for example, the researchers make sure they are truthful while interpreting the research data); fairness (focuses on ensuring the researcher is fair to all participants and their rights are upheld).

The ethical considerations undertaken in this study will be discussed in a separate section. However, it can be assured that the researcher endeavoured to be as sound as possible in maintaining the four criteria mentioned above. Ethical approval was sought from the University of Melbourne for this study and guidelines provided by the Australian Research Council (ARC) were followed.

3.1.4 Pragmatism as the research paradigm of this study

After considering the ontological and epistemological assumptions behind different research paradigms, it was decided that this research will adopt pragmatism as its research paradigm. Cohen et al. (2018) suggest that in educational research, foundationalism and quest for absolute knowledge is questionable. Discussing various paradigms and approaches to research, they argue for methodological, paradigmatic and theoretical pluralism. Pragmatist researchers consider reality to be both singular (e.g., a theory may exist to explain the phenomenon of interest) and multiple (e.g., it is important to also consider a range of individual inputs to explain the phenomenon)

(Creswell & Clark, 2018). Adopting pragmatism makes it possible for researchers to be eclectic while choosing methods (Denscombe, 2010). Instead of arising from the antecedent conditions (as in post-positivism), pragmatism stems from actions, situations and consequences and focuses attention on the research problem using pluralistic approaches in deriving knowledge about it (Creswell & Poth, 2018). Based on their review of other researchers' work, Kivunja and Kuyini (2017) outlined some characteristics of the pragmatist paradigm in their paper. Among them the following were deemed beneficial for adopting this worldview in this study:

- a. Rejecting the positivist idea that social science inquiry can be used to uncover the 'truth' about the world and rejecting the forced choice of locating a research within either a positivist (post-positivist) or an interpretivist (constructivist) paradigm.

- b. Emphasising 'workability' in research, using 'what works' – allowing researchers to focus on the question(s) they are exploring instead of spending time on whether the questions are exclusively quantitative or qualitative.

- c. Adopting a worldview that makes it possible to use research design and methodologies best suited for the purpose of the study; employing lines of action best suited for researching the phenomenon under consideration.

- d. Searching for useful points of connection in the research project that will be beneficial in understanding the situation (adapted from Kivunja & Kuyini, 2017, pp. 35 - 36).

The ontological beliefs, epistemological beliefs, axiological beliefs and methodological beliefs of pragmatism are as follows (Creswell & Poth, 2018; Kivunja & Kuyini, 2017):

Ontology: There is no single reality, individuals make their own unique interpretation of reality based on what is useful, what works and what is practical.

Epistemology: reality is discerned by utilising various research tools generating both deductive (objective) and inductive (subjective) evidence, it is relational, and the researcher decides what relation is suitable for a specific research.

Axiology: doing research that is beneficial for people, and discussion of values is essential because of the way both the researchers' and participants' views are reflected in the knowledge.

Methodology: Combination of quantitative and qualitative approaches of data collection and analysis, that is, a mixed methods methodology.

The following section discusses the methodology chosen based on the decision of adopting pragmatism as the research paradigm.

3.2 Research Methodology

Methodology is the processes and procedures used in a research study (Creswell & Poth, 2018). Methodology for a research study is decided based on the ontological and epistemological position of a researcher (Tuli, 2010). It outlines the principles, processes and practices of research that the researcher intends to follow. According to Scotland (2012), methodology revolves around the why, what, from where, when, and how data is gathered and analysed. It depicts the logic and flow of the research process and includes the finer details of the research project (e.g., assumptions made,

limitations faced and how they were tackled, decisions on data gathering instruments, selection of participants, collection and analysis of data, etc.) (Kivunja & Kuyini, 2017). Specific techniques and procedures used in collecting and analysing data are known as methods. These can be “traced back, through methodology and epistemology to an ontological position” (Scotland, 2012, p. 10).

Creswell and Creswell (2018) defined research approaches as “plans and procedures for research that span from broad assumptions to detailed methods of data collection, analysis and interpretation” (p. 40). These approaches are synonymous with research methodology. They discuss three research approaches: qualitative, quantitative and mixed methods. They also explain that these are not rigid categories or dichotomies. The choice of research methodology depends on the researcher. Given the option of choosing among quantitative, qualitative and mixed methods approaches, the choice is usually made based on his/her preferred worldview/paradigm/research philosophy. In addition, this decision depends on the research problem, the personal experience of the researcher and the audience for the research (Creswell & Creswell, 2018). As mentioned earlier, this research adopted pragmatism as its worldview. Pragmatism provides a set of assumptions about knowledge and enquiry that is suitable for the use of a range of methods and research strategies and is considered to be the natural philosophical partner of mixed methods approach (Denscombe, 2008, 2010).

The rationale behind choosing a mixed methods approach was that in order to answer the research questions of this study, quantitative or qualitative research alone would not be adequate, and integrating the strengths of both could provide the best

understanding. In addition, the researcher's previous experiences and training also came into play while deciding on a mixed methods approach as he has had previous experiences of working in quantitative and qualitative studies and has done some mixed methods research as well. Although the time and resources necessary for conducting a mixed methods research study were quite demanding, this seemed to be the best research approach for adequately addressing the research questions. Furthermore, consideration of the audience for this research also supported choosing a mixed methods approach. Creswell and Creswell (2018) mention that while selecting research approaches students ought to consider the approaches supported by their advisers. The advisory committee for this study acknowledged the necessity of using multiple data sources and using both quantitative and qualitative data. Overall, based on the research problem, personal research experience, audience for this study and research philosophy, a mixed methods approach was selected.

3.2.1 Mixed methods as a research methodology

In recent years mixed methods research has advanced rapidly, though it has long been used without naming it as a paradigm (Denscombe, 2008). The core assumption of mixed methods approach is that integrating qualitative and quantitative data generates additional insights beyond the information gathered from either of them singularly (Creswell & Creswell, 2018). In collecting empirical evidence to validate the hypothesised theoretical framework for assessing arithmetic word problem solving in this research, multiple sources of validity evidence were deemed suitable. Some of these are quantitative (e.g., evidence based on test content, relation with other

variables, etc.), whereas others are qualitative (e.g., evidence based on response processes, i.e., cognitive interviews). Based on discussions of different aspects of mixed methods research in literature (e.g., Denscombe, 2008, 2010, 2014), these are the characteristics of mixed methods research that were deemed beneficial for this study:

- A variety of perspectives for viewing the research problem providing a more rounded description/ more complete explanation of the phenomena studied rather than a single approach, and ensuring accuracy of findings;
- Combining different types of research (qualitative and quantitative) in one project;
- Choosing methods based on “what works best” in tackling specific problems;
- Clearly explaining the sequence and priority of quantitative and qualitative elements and how they link with one another.

Despite the benefits, using a mixed methods approach also contains some challenges that should be kept in mind. As Denscombe (2014) outlines, using a mixed methods approach makes extra demands on the researcher, as to do this a researcher must have skills in both quantitative and qualitative research. As data analysis entails running different types of analysis and comparing, contrasting and integrating the findings from them, it can get quite challenging. Before deciding to use a mixed methods approach, the researcher should consider these factors and decide what is best for the research project in terms of time, expertise and other resources.

There are different perspectives within the mixed methods approach on how quantitative and qualitative elements should be utilised within a single project and

whether to integrate them, combine them or use them in tandem (Denscombe, 2008).

Based on these perspectives there are many different designs that can be used for a mixed methods research. Creswell and Creswell (2018) discuss three primary designs of mixed methods research found in the social and health sciences research:

- a. *Convergent mixed methods*: quantitative and qualitative data collected almost concurrently integrated to develop a complete analysis of the research problem explaining contradictory or incongruent findings (if needed).
- b. *Explanatory sequential mixed methods*: starts with collection and analysis of quantitative data, then results from quantitative analysis are further explained through collection and analysis of qualitative data.
- c. *Exploratory sequential mixed methods*: starts with qualitative data collection and analysis in order to build a knowledge base for future quantitative research

Among these the convergent mixed methods design was considered to be suitable for this study. Quantitative and qualitative data were collected simultaneously. These were later integrated to develop an analysis where findings from the different instruments complemented one another. The following sections discuss the different aspects of data collection and analysis.

3.3 Data collection instruments

Wilson (2005) defines instruments as techniques used for relating observations of the real world to measuring something extant as part of a theory (latent or unobserved). Several instruments were used in this study to establish the validity and reliability of

the theoretical framework developed for Arithmetic Word Problem Solving. Table 3.1 outlines the different types of validity evidence (American Educational Research Association [AERA] et al., 2014) collected through different measures.

Table 3.1
Sources of Validity Evidence

Type of Evidence	Source
Evidence based on test content	Test design and specification while designing the test Expert opinionnaire
Evidence based on response processes	Cognitive interview of respondents
Evidence based on internal structure	Rasch analysis of the test (the Wright map, item fit, response fit and so on)
Evidence based on relations with other variables	Textbook AWPS items in the test Mathematics and Bangla test scores from school tests
Evidence based on consequences of testing	Not suitable

Developing a test based on the proposed AWPS framework formed the core of instrument development. In addition, other methods were employed to collect pieces of various validity evidence. These included integrating textbook items with the items developed based on the AWPS framework within the student test, using cognitive interviews (think aloud protocols), collecting expert opinion regarding the AWPS items (expert opinionnaire), collecting students' literacy and numeracy scores from their school-administered tests, and collecting school background information (including information about the mathematics teachers) through a school questionnaire. This

section discusses how the different data collection instruments used in this research were developed.

3.3.1 Student test

A well-designed instrument plays a crucial role in the process of psycho-social measurement (Wu et al., 2016). Using a detailed theoretical specification of the construct that is being measured helps in supporting the validity of the inferences and decisions that are made based on the measures (Wolfe & Smith, 2007). In other words, following predefined and research supported procedures while designing an instrument is the foundation for the validity argument; this is the evidence based on test content. Various authors have discussed the systematic process of developing a test for measuring a construct and simultaneously documenting evidence for validation of the instrument. Wilson (2005) outlines four building blocks, which are used as steps in an iterative process to develop a psycho-social measurement. These building blocks are (i) the construct map (identifying and defining the construct), (ii) the item design (developing items), (iii) the outcome space (gathering and scoring item responses) and (iv) the measurement model (producing measures). The steps of test development discussed by Wolfe and Smith (2007) build on this model and they further unpack the four building blocks, suggesting the development of a test should involve defining the purpose of the test, formulating the test specifications, developing items, doing expert reviews and conducting pilot studies. Steps of test development outlined by Wu et al. (2016) are similar to Wilson's (2005) four building blocks. The steps they have discussed are – defining the construct, writing, administering and scoring test items, and producing measures. In this study, the four

building blocks model of Wilson (2005) was followed as the key concept. In addition, the other references helped in clarifying the concepts discussed in various steps of developing a test.

The following sections discuss how the first two building blocks of the model, that is, identifying the construct and developing items were utilised in this research. How the other two building blocks were utilised is discussed in later sections (in the data analysis methodology sections). Activities related to designing the students' test in this research spanned across the pilot and the final data collection. The first set of items developed through the processes discussed here went through a second iteration of these activities where items piloted were revised based on the empirical data and prepared for the final data collection. Hence, the discussion of these activities is not separately presented under piloting and final data collection.

3.3.1.1 Identifying and defining the construct

Before designing an instrument, it is essential to clarify the construct or latent trait that is being measured (Wu et al., 2016). It is also important to outline the purpose for which the instrument will be used (Wolfe & Smith, 2007). In this study the construct was part of a theoretical model of a person's cognition. The purpose of developing the measures was to compare the theory-based predictions with empirical observations, that is, checking if the postulated theoretical framework of Arithmetic Word Problem Solving is acceptable in terms of validity and reliability. Another purpose of developing the measures was to assess the student's achievement in the construct of interest so their achievements may be compared with other existing measures. Measures can be developed to be used for more than one purpose (Wolfe & Smith, 2007).

The construct of interest in this study is arithmetic word problem solving. As discussed in the previous chapter, five areas within this construct were identified to define the construct. These areas were identified based on a review of literature and more literature reviewing was conducted to better understand ways of assessing these. After clarifying the nature of the construct, a researcher has to determine the types of observations of behaviour, products, and actions that can help in eliciting evidence for making inferences about the construct (Wilson, 2002). This comes in the form of an instrument for data collection.

While developing an instrument based on a construct, it is helpful to communicate the construct through a construct map. A construct map assumes that the construct of interest extends from one extreme to another (e.g., small to large, high to low or simple to complex) and that there are qualitatively distinguished levels in between (Wilson, 2005). A construct may be unidimensional or multidimensional. A construct is unidimensional when there is a single latent variable that underlies the construct. On the other hand, a construct is multidimensional when it is formed with multiple latent variables. For this research, arithmetic word problem solving is considered as the construct. In current practices of assessing arithmetic word problem solving, it is considered as unidimensional (e.g., as discussed in Chapter 1 of this thesis, the teacher editions in Bangladesh only focus on the word problem as a whole). However, it is postulated in this research that this construct has multiple dimensions (as outlined in the theoretical framework of AWPS). Although AWPS is regarded as a multidimensional construct, the four building blocks model can still be used by developing a construct map for each dimension (Wilson, 2005). These

dimensions or elements of arithmetic word problem solving are also called strands in this research.

There are two approaches to developing a construct map. Keeping in mind the idea that there is a qualitatively distinct set of continuous categories that goes from more of something to less of something (or more correct to less correct; or complex to simple, etc.), the construct map may reflect respondents' position on the construct (from more able to less able) or it may reflect the position of their responses (more difficult to less difficult or complex to simple). Both are eventually needed but for the sake of developing a test, either has to be selected as a starting point. For this research, the qualitatively different responses were considered as the starting point and for each of the strands of AWPS one such map was postulated. Wolfe and Smith (2007) describe a concept similar to construct maps which they call construct modelling. They discuss four different construct models: internal model, external model, developmental model and processing model. Among these, the developmental model seemed suitable for conceptualising the construct maps used in this study. In a developmental model, the manifestation of a construct changes as an individual moves from novice to expert state (or under-developed to mature state, etc.). These states are defined based on theoretical assumptions about what it means to have high, medium or low levels of proficiency in the domain of interest, which in turn helps in item writing. There are a number of sources of information that help in building construct models, including real-world observations, literature review on the construct and/or related empirical research, review of previous instruments, expert and lay perspective of the construct, and so on (Wolfe & Smith, 2007).

The process for specifying the ability levels within the construct map had three steps. First, the skills and concepts relevant for a strand of the hypothesised AWPS framework were identified. Second, a set of skills suitable for the year levels targeted in this research (grade 4 and grade 5) were selected. Finally, the selected skills were classified from easy to difficult. The first step was done based on review of literature relevant to the strand. The second and third steps were based on review of mathematics curriculum, textbooks, test items and expert input.

In order to develop the construct maps (and later write items reflecting the construct maps), this research tapped into a number of different resources. This included a review of the mathematics curriculum, mathematics textbooks, theoretical papers and empirical papers related to the postulated strands of arithmetic word problems solving; review of existing test items; and expert review. It is difficult to point out which references were used for what decisions as the review provided a general idea of how to articulate and assess each of the strands and some references were resources for more than one strand. However, the following section discusses some examples of the reviewed literature for the different strands.

Literature discussing reading comprehension in general (e.g., Deane, Sheehan, Sabatini, Futagi, & Kostin, 2006; Duke & Pearson, 2009; Fletcher, 2006; Hudson, Lane, & Pullen, 2005; Kintsch, 1998) and reading comprehension in mathematics (e.g., Adams, 2003; Bergqvist & Österholm, 2010; Kintsch & Greeno, 1985; Österholm, 2006) was consulted to better understand comprehension of the problem and to develop a construct map. A number of concepts were outlined by these authors. For example, Deane et al. (2006) outline how effective reading entails more than lexical

sets or vocabulary and depends on understanding of sentence structure and combining information from multiple sentences. They identify sentence complexity (for example use of connecting words such as *since*, *while* and use of temporal subordinators such as *after*, *before*) as an important factor for difficulty in reading comprehension and highlight how text complexity increase developmentally. Bergqvist & Österholm (2010) suggest that the implicit model of the reading comprehension of a mathematical task is a function of two variables – prior knowledge and text complexity. According to them “reading comprehension increases with better prior knowledge and with lower text complexity” (P. 49). Kintsch (1998) distinguishes three different components of mental representation of texts – the surface component (the words and phrases), the textbase (the semantic structure of the text), and the situation model (consolidation of the textbase and readers’ prior knowledge). In addition to this literature, mathematics curriculum and textbooks were reviewed while drafting the ability levels for the comprehension of the problem strand. Then the levels were classified from low to high based on expert input.

Riley et al. (1984) provide a typology of addition and subtraction word problems (change, combine, compare, and equalising problems and their sub-sets). Other typologies of addition and subtraction word problems were reviewed in Willis and Fuson (1988), Verschaffel and De Corte (1997) and Thevenot and Barrouillet (2015). These typologies and Fuson’s (1992) chapter with a list of additive and subtractive situations were consulted for developing the ability levels in the second strand, that is, ‘mapping conceptual relations to a numerical representation’. It should be noted that, the typologies of problems reviewed in this literature provided some background for articulating the levels in the third and fourth strand too.) In addition,

Wu's chapter on translation of verbal information into symbols (Wu, 2016); Vergnaud's (1982) "classification of cognitive tasks and operations of thought involved in addition and subtraction problem" (p. 39), and the model of representation discussed in Kintsch and Greeno's (1985) article were helpful for determining the concepts relevant to the second strand. After identifying the concepts relevant to the strand, the ability levels were defined based on the curriculum, textbooks, and expert input.

The third strand - 'selecting an appropriate strategy for solution' was the most complicated to define and assess. Literature reviewed for this research, where strategy in the context of mathematical problem solving is discussed, can be classified into two sets. This is an over-simplified classification of the ways strategies have been discussed but helpful for this research. One set discusses strategies in a broad sense where strategies are regarded as procedures for solving the entire problem (e.g., Ashcraft, 1990; Clements, 1980; Rieskamp & Otto, 2006; Siegler, 1999; Willatts, 1990). Some discuss typologies of strategies such as forward search and problem reduction (Willatts, 1990); backward thinking and forward thinking (Sweller, 1988). The other set of literature considers strategies as techniques for calculation where the concepts revolve around mental computation and mental arithmetic (e.g., Campbell & Xue, 2001; Carr & Davis, 2001; Fennema, Carpenter, Jacobs, Franke, & Levi, 1998; Gravemeijer & Stephan, 2002; Hartnett, 2007, 2008; Ineson, 2007; Lemaire & Arnaud, 2008; McIntosh, 2002; Wright, Martland, & Stafford, 2006). These concepts of strategy vary broadly. For example, Fennema et al. (1998) categorise strategies into a) modelling or counting, b) invented algorithms, and c) standard algorithms; Campbell and Xue (2001) classify strategies into transform, count, retrieval and other; Carr and

Davis (2001) describe retrieval, manipulation, counting-all, counting back as strategies; Lemaire and Arnaud (2008) identify nine strategies used by their respondents for addition of two two-digit numbers –“rounding the first operand down, rounding the second operand down, rounding both operands down, columnar retrieval, rounding the first operand up, rounding the second operand up, rounding both operands up, borrowing units, and direct retrieval” (p. 9); and Hartnett (2007) classifies mental computation strategies into a number of categories (e.g., “count on to add, count back to subtract, count on to subtract” p. 349, etc.) and links them to literature. It is noticeable how some of these authors have gone into more detailed classification of strategies while others have grouped them together. The theoretical underpinnings of this research aligned with the conception of strategies as mental techniques of calculation. Elaborate examples and extensive lists of mental strategies for addition and subtraction were found in Bradley (2006). However, review of curriculum and textbooks from Bangladesh showed that students in Bangladeshi schools are not explicitly taught about these strategies. In addition, after looking at the items that can be used to elicit evidence on such mental strategy use, it was felt that it will be difficult to adequately assess this type of strategy within a broader test of AWPS given the time constraints in administering a test. In addition, Bangladeshi students and teachers might not be familiar with such strategies or might face difficulty in communicating these, even if they were implicitly using them. To alleviate these concerns, another type of strategy was considered for inclusion in this test. These are much simpler and focus on one particular aspect of calculating. Elementary students have to use carrying and borrowing for addition and subtraction and this is an important part of what teachers have to teach them in schools. So, it was decided

that carrying and borrowing strategies will be assessed through the AWPS test. In the pilot test, items focusing on assessing the more elaborate strategies of addition and subtraction (i.e., counting up, counting down, and such) were included in addition to items tapping into the carrying and borrowing strategies in order to check students' familiarity and comfort with the former. However, initially it was difficult to identify a gradation of demands for the items assessing carrying and borrowing. This was done in terms of the number of separate steps necessary in the solution process and adding on more steps was postulated to increase the cognitive load for solving the item, hence making it more difficult. Eventually the items reflecting mental computation strategies had to be removed in the final test, as the students faced trouble in answering and were quite uncomfortable with those (this will be further discussed in the revision of items in the pilot study chapter). However, some evidence of students' mental calculation strategy use was elicited in the cognitive interviews. In addition, the number of digits were increased in the final version of the test (compared to the pilot test) to increase the difficulty based on advice from experts. This was also supported by literature (e.g. Willis & Fuson, 1988).

Developing the construct map and items for strands 4 (performing the calculations) and 5 (verifying the solution and communicating the results) was largely informed by the review of the Bangladeshi primary curriculum (National Curriculum and Textbook Board [NCTM], 2011) and the mathematics textbooks of primary level there (Khan, Matin, Begum, & Uddin, 2013; Mollah, Ullah, Halder, & Dhali, 2013a, 2013b, 2013c, 2013d; NCTB, 2009). Review of the curriculum and textbooks from Bangladesh, the TIMSS and PISA assessment frameworks and other publications related to these studies (e.g., Mullis & Martin, 2013; Stacey & Turner, 2015) and

review of items from the SMART tests (Stacey, Price, Gvozdenko, & Steinle, 2013) helped in the articulation and item development of all the strands.

Professionals experienced in developing instruments in a relevant area are considered as a great resource for developing items (Wilson, 2005). They can help in identifying specific issues with items and how information gathered at different stages of item development can be integrated. Experts from the Melbourne Graduate School of Education (MGSE), including the supervisory panel for this research, and other academics from the mathematics education group and primary mathematics teachers from Bangladesh played a key role in the development of the construct maps, item specifications and the items by providing thorough advice and hands-on reviews during the drafting, revision and finalisation. The construct maps for piloting and final data collection are included in Appendices 3.1 and 3.2.

3.3.1.2 Developing items

The second building block is identifying a way of manifesting the theoretical construct into a real-world situation, that is, developing items that are realisations of the construct (Wilson, 2005).

At this stage, causality is often unclear. The construct may cause the item responses, or the items might cause the development of the construct, and on many occasions clearly defining a construct is dependent on developing and trying out a large set of items (Wilson, 2005). The item development process for this research took an iterative form (both generally and within the item development stage). The item writing phase started with a tentative understanding of the five strands. As the item

construction activity continued, the strands of the construct also became better-defined.

There are many types of items that can be used. Items can broadly be divided into two categories: selected response items and constructed response items (Haladyna & Rodriguez, 2013). Multiple choice items, a popular type of selected response item, are commonly used in achievement testing. This is a forced-choice type of item, as respondents have to choose from the limited range of alternatives provided to them. Discussing the pros and cons of multiple-choice items, Wu et al. (2016) commented that despite some limitations, these are cost-effective in providing data for measurement.

It is crucial to understand the relationship between an item and a construct. Due to various constraints, a researcher has to choose a finite set of items from the large number of items that could possibly be used for measuring the construct (Wilson, 2005). In order to select a suitable set of items, it is important to first develop a large pool of items that reflect the construct and then devise a well-defined plan for selecting a relevant and representative subset of items from the pool that is practical for administration. Using a test specification might be helpful in this selection process (Wolfe & Smith, 2007). A test specification or test blueprint contains the complete operational definition of all test characteristics (item format, number of items, item allocation to different areas, etc.) (Downing, 2006).

Test specifications are established considering the purpose of the test, student characteristics, content coverage, duration of assessment and other constraining factors (Wu et al., 2016). Following a well-informed test specification to develop items

helps in developing validation arguments regarding the substantial aspect of validity, that is, validity evidence based on test content, as theory-based predictions about item difficulty can be compared with empirical difficulties (Downing, 2006; Wolfe & Smith, 2007). The test specifications for the pilot and final test are included in Appendices 3.1 and 3.2 along with the construct maps.

According to Wilson (2005) there are two components in item design, the construct component and the descriptive components. The construct component deals with the representation of the construct in the items so items can provide a criterion-referenced interpretation along the construct and the descriptive components represent all the other characteristics that a set of items needs to have (e.g., item format, response types, etc.). Once the construct map was used to decide on items, the other decisions were part of the descriptive component of item design for this research. It was decided that a multiple-choice format will be used for most of the strands where each item will have five alternatives. One of the strands (selecting strategy) will have items in open-response format. Other descriptive components of item design included deciding the content level appropriate for the respondents, ensuring reflection of the curriculum in the test, and familiar content and format for the respondents.

After establishing the construct maps and delineating the test specifications, a large item pool was developed where at least ten items were written for each of the strands. This item pool was used for item panelling (where items were reviewed and revised by a group of experts) and a draft set of 34 items was chosen for piloting. Including the textbook AWPS items, the pilot test booklet consisted of 40 items. Wu et

al. (2016) suggest that “if the purpose of the test is to determine whether a student has an average, above average or below average performance in mathematics, a 40-item test (or a 40 score-point test) will be sufficient to provide that information” (p. 31). Forty items also seemed suitable to be administered within a limited time in the schools. Based on empirical data collected in piloting, the 34 items written were revised and 32 were retained for the final data collection. The six textbook items were retained and so the finally administered test consisted of 38 items in total.

3.3.1.3 Selection of textbook items

Six items were selected from Bangladeshi primary mathematics textbooks to represent arithmetic word problems. This was done to collect validity evidence based on relations with other variables. Table 3.2 outlines the sources of these items. These were complete arithmetic word problems from addition and subtraction related chapters within the textbooks. They were made part of the test booklet to simultaneously collect data that can be used to relate students’ performance in the postulated theoretical framework with an existing type of assessment.

Table 3.2

Sources of Textbook Items Included in the Test Booklet

Item Number in test	Textbook Grade level	Source	Chapter name
6	3	(National Curriculum and Textbook Board [NCTB], 2015b)	Relationship between addition and subtraction
12	3	(National Curriculum and Textbook Board [NCTB], 2015b)	Relationship between addition and subtraction

18	4	(National Curriculum and Textbook Board [NCTB], 2015a)	Addition and subtraction
24	3	(National Curriculum and Textbook Board [NCTB], 2015b)	Relationship between addition and subtraction
30	4	(National Curriculum and Textbook Board [NCTB], 2015a)	Addition and subtraction
36	4	(National Curriculum and Textbook Board [NCTB], 2015a)	Addition and subtraction

3.3.1.4 Construction of test booklets

It is important to consider instrument format while organising items into a test. An instrument or test can consist of a single item format (e.g., just multiple-choice items) or a more complex mix of different item formats (e.g., multiple-choice and open-ended) (Wilson, 2005). Item format selection is usually determined by the limitations and benefits inherent in various formats. The test constructed here contained mostly multiple-choice items, with a few open-ended items. The multiple-choice format is easy to score objectively and can be used to assess a wide range of content within a short time (Wolfe & Smith, 2007). However, some open-ended items were included for the third strand of the framework (selecting an appropriate strategy for solution) because of the difficulty in assessing the content through a multiple-choice format. Some multiple-choice items were included (two) in the pilot test for this strand but based on the empirical evidence collected in the pilot study, only multiple-choice items were included in the final test.

While putting together a test from items, it is suggested that the test starts with easier items and gradually moves on to items of higher difficulty to ease the

students into the test and ease any anxiety the test taker might have, or to at least delay it (Wu et al., 2016). The test was constructed keeping this suggestion in mind and started with the easiest items and moved on to more difficult items gradually. Items from the different strands were presented in a cyclic order along with the textbook items. So, item 1 was from strand 1, item 5 from strand 5, item 6 a textbook item and then item 7 from strand 1, and so on.

In addition to the items, the test had some questions that collected background information (such as gender, age, etc.) about the students. These were included in order to conduct analyses based on demographic variables. It was envisaged that the demographic variables could be useful in conducting comparative analyses between students of different backgrounds. There were two separate test booklets for the pilot study and the final study. The pilot study test booklet can be found in Appendices 3.3 (English) and 3.4 (Bangla). The answer key for the pilot study test is included in Appendix 3.5. The final test booklet can be found in Appendices 3.6 (English) and 3.7 (Bangla). The answer key for the test conducted in the final phase is included in Appendix 3.8. The Bangla translation of the test booklet was administered in both iterations of data collection.

3.3.2 Students' cognitive interviews

Including a data collection method for eliciting respondents' voices regarding how they responded to the items constitutes an important part of developing items (i.e., the second building block of Wilson's (2005) model). In addition to the formative use of such evidence for item design, they can be used as a summative tool for gathering validity evidence (Wilson, 2005).

Ericsson and Simon (1980) first posited the idea of using verbal reports as a source of information in research. They suggested that the notion that verbalisation only provides casual information which needs to be verified by other data was wrong, and this has had a substantial impact on the ways in which verbalisation data are collected and analysed. Given verbalisation is used to mainly generate ideas and hypotheses, according to them, many investigators have disregarded reporting on the methodologies used for data gathering and data analysing (which vary vastly) using verbal reports. They attended to these gaps in the literature and, based on the theoretical framework of human information processing, outlined how verbal data can be used as a source of information.

As one of the sources of validity evidence, verbal reports have been beneficial for researchers who wanted to validate the response processes elicited by test items within large-scale assessments and performance-based assessments (Leighton, Heffernan, Cor, Gokiart, & Cui, 2011).

The 1999 Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA] and National Council on Measurement in Education [NCME], 1999) was the first to explicitly consider the study of the response processes to test items and questionnaires as a source of validity evidence. The 2014 Standards also discuss response processes as a source of validity evidence (American Educational Research Association [AERA] et al., 2014). Padilla and Benitez (2014) grouped the methods of obtaining validity evidence through the response processes into two categories: i) methods that “directly access the psychological processes or cognitive operations

(thinkaloud, focus group, and interviews)” (p. 139)) and ii) methods that “provide indirect indicators which in turn require additional interference (eye tracking and response times)” (p. 139). This study focused on directly accessing the psychological processes through cognitive interviews. However, such a classification can be too simplistic. Wilson (2005) discusses two major types of investigations that tap into psychological processes through eliciting data about responses: think-alouds (where students are asked to talk about their thinking while responding to the items, aka cognitive labs) and exit interviews (where students verbalise their thinking after responding to the items, might be after each item or at the end of the complete test). Wilson (2005) further comments that the data collection strategy that involves both think-alouds and exit interviews might be best.

Beatty and Willis (2007) described cognitive interviews as “the administration of draft survey questions while collecting additional verbal information about the survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends” (p. 288). In contrast with ordinary interviews, which focus on generating codable responses to the questions, cognitive interviews “focus on providing a view of the processes elicited by the questions” (Presser et al., 2004, p. 4). Information from respondents collected through such interviews is important as it gives the researcher an opportunity to reflect on and decide whether changes are necessary in the construct, specific items or item types, outcome space or scoring schemes (Wilson, 2005).

There are two types of objectives for conducting cognitive interviews: formative (which Willis, 2015 calls reparative) and confirmatory (which Willis, 2015 calls descriptive). In the reparative approach, interviews are conducted to inspect and repair items and in the descriptive approach interviews are done in order to collect evidence on how the items function as measures of the intended concepts. The latter is done as part of collecting validity evidence. This study adopted the descriptive approach for the final phase of data collection as the intention was to gather validity evidence on how the items worked as measures of the concepts they were designed to assess.

From the review of the literature it was apparent that there are a few different names given to the same or similar approaches of data collection techniques which focus on collecting data regarding what goes on in a respondents' mind while answering an item. The names include cognitive interviews, cognitive labs, think aloud protocols, and so on. Though titled differently, they have more similarities than differences. Based on the review of existing practices (e.g., Collins, 2003; Ericsson & Simon, 1980; Gadermann, Guhn, & Zumbo, 2011; Johnstone, Bottsford-Miller, & Thompson, 2006; Leighton et al., 2011; Linninger et al., 2015; Miller, 2014; Miller, Willson, Chepp, & Ryan, 2014; Padilla & Benitez, 2014; Presser et al., 2004; Willis, 2004, 2015) it was felt that it is better to conceptualise the whole idea as a "cognitive interview" where "think aloud" is a part of it. This think-aloud can be concurrent (when a respondent is verbalising while answering the item) or retrospective (when verbalising a mental process that occurred at an earlier point of time). Furthermore, the think aloud may or may not have verbal probing (where the respondents are asked probe questions about their responses). If it is decided that probing will be

used, it can also be concurrent or retrospective. In concurrent probing respondents are asked questions immediately after answering an item whereas, in retrospective probing, they are asked questions after they have gone through all the items. After considering the different options it was decided that a concurrent think-aloud with some retrospective verbal probing will be used in this study. However, the concurrent think aloud was in a sense retrospective because the participants had already sat for the test before coming to the interview. Having said that, their responses were taken while they answered the items for a second time, so in that sense they were concurrent think-alouds. In addition, they were asked to clarify some responses sometimes after they had performed the think-aloud for a particular item. These 'probes' were given right after they had answered an item. So, they could be considered as concurrent probing.

The data collection instrument for the cognitive interview consisted of the items from the test. Respondents were asked to verbally articulate their thinking while solving the items. In addition to the items there were some questions that tapped into respondents' use of strategy. Furthermore, respondents were asked which of the items they felt was too difficult or made them feel uncomfortable. Measurement of human traits is different from measuring a physical property of a matter. It is important to determine if instruments used in research are in anyway offending or harmful to the respondents. This is one of the reasons why a set of items should be piloted with a small number of respondents before large scale administration. Wilson (2005) argues that in a think-aloud/ cognitive interview, there should be specific prompts that elicit respondents' discomfort or distress in answering items.

3.3.3 Expert opinionnaire

The expert opinionnaire was constructed based on the student test items. There are a number of different ways for collecting experts' input about test items to elicit validity evidence. Some researchers (e.g., Kamei, 2019; Strickland, 2018) have used workshops and interviews. However, due to the constraints of this study, it was decided that an opinionnaire will be used, which makes it possible to collect data even without the presence of the researcher. The opinionnaire was quite straightforward. It had some background questions about the expert in the beginning to collect demographic information. Next, the purpose of the opinionnaire was explained and the experts given a brief introduction to the theoretical framework of arithmetic word problem solving with the five strands. Then the test items were presented, and they were asked to select which strand they thought each of the items were best suited for. One item was done as an example so they could understand what they are supposed to do. Researchers such as Crocker and Algina (2008) have highlighted using such methods for collecting experts' opinion about content related validity evidence of a test. The English version of the expert opinionnaire can be found in Appendix 3.9 and the Bangla version (which was administered) can be found in Appendix 3.10.

3.3.4 Students' Mathematics and Bangla test scores from the school

As the research focused on determining the validity of the postulated theoretical framework of arithmetic word problem solving, different validity evidence types were collected. In order to collect evidence based on relations with other variables, AWPS items from the textbook were included within the students' test. Furthermore, participating students' test scores from their school Mathematics and Bangla

examinations were collected so their performance in those tests can be compared to their performances in the test administered in this research. The schools in Bangladesh administer three school-level examinations in an academic year. The student data collection for this research occurred just after the mid-year examination and all the schools that were part of this study had the same set of items for the Mathematics and Bangla as they were under the same regional administration. Though the test administered for this research was targeted at the postulated five strands of AWPS and the school tests focused on overall Mathematics and Bangla (language) competency, it was hypothesised that there should be positive correlation between them. Wu et al. (2016) suggest that “if a test consists of problem solving items where each item requires the same ‘proportional amount’ of reading ability and mathematics ability, the test can still be considered ‘unidimensional’, with a single latent variable called ‘problem solving’ ”(p. 24, emphases in original). The construct of problem solving is a composite of abilities in this case.

3.3.5 School background questionnaire

A questionnaire was developed to collect demographic data on the participating schools. This enabled the researcher to collect data about the location of the school and the background of the teachers who taught mathematics in grades 4 and 5. The school background questionnaire can be found in Appendix 3.12 (English) and Appendix 3.13 (Bangla).

3.3.6 Translation of instruments

The medium of instruction in majority of the Bangladeshi primary schools are Bengali (endonym Bangla). So, it was necessary that the tests administered with the students

and students' post-test interviews are translated into Bengali. Even though some of the experts would be comfortable using English for the expert opinionnaires, it was also translated to Bengali to ensure language barriers did not inhibit the expression of opinions by experts who are not confident users of the English language. These instruments were initially developed in English and then translated into Bengali. However, since it was known that these would be translated into Bengali, every effort was made at the item writing stage to include straightforward instructions, stems and alternatives.

Literature suggests that there are two main methods of translation for research instruments: direct (one-way translation) and back translation (Hambleton, 1993; McKay et al., 1996). In direct translation, a bilingual person translates the source (original) language instrument into the target (second) language. Direct translation or forward translation is the simplest method for instrument translation and the most widely used in cross-cultural research (Phongphanngam & Lach, 2019). A variation of direct translation method, 'translation by committee' was used for the instruments in this study. Based on objectives, translations can be literal, conceptual or culturally equivalent. Conceptual translation is highly recommended as, even in the least resource intensive method (one-way translation), it was found to be effective in generating comparable translations (McKay et al., 1996). Hence conceptual translation was set as the objective in this study.

Committee translation, also known as parallel translation or team translation, involves several translators and a translation coordinator, all of whom are skilled in the source and the target languages, competent on the subject material of the

instrument and in the skills required in implementing the instrument. They separately translate the source instrument into the target language and then bring their translations to a reconciliation meeting.

In the meeting, the translation coordinator along with the translators evaluate the different versions of the translations. They resolve any extant discrepancies and finally agree on a version that includes the best of the independent translations or what emerges from the reconciliation meeting discussions (Harkness & Schoua-Glusberg, 1998; McKay et al., 1996). Large scale international assessment such as TIMSS have also used a similar method for instrument translation (Martin, Mullis, & Chrostowski, 2004, p. 93).

Since the translations were not of established instruments and it was possible to change the source language instrument, a *decentring* strategy was used along with committee translation. Beginning with a draft instrument in the source language, decentring produces a final instrument in two languages (source and target) through a process of translation and paraphrasing between the two languages (Harkness & Schoua-Glusberg, 1998). In this strategy, when the translation process identifies words or phrases in the source instrument that are deemed problematic in the target language, both source and target instruments are amended until comparable questions or items are attained and both source and target instruments are finalised concurrently (McKay et al., 1996).

Two former colleagues of the researcher, one currently working at the National Academy of Primary Education (NAPE) and another working as an Instructor at a Primary Teachers Training Institute (PTI) in Bangladesh helped with the

translations. They were selected for their skill in the source and target languages, that is, English and Bengali, their familiarity with mathematics education in the primary level and their experiences of educational research. The researcher worked as the translation coordinator. Three independent translations were made for each instrument and the final version was decided through discussion and corroboration at reconciliation meetings. However, after revising the student data collection instruments (test and interview) based on piloting, the final version of the test was translated in collaboration with only one of the translators (the one working at NAPE) since the other translator was not available.

Most of the changes coming from the translation reconciliation came for the items in the student tests. The student interview and expert opinionnaire were based on these test items so these changes were implemented in those instruments too. A list of changes made to items can be found in Appendix 3.14.

3.4 Participants of the study

This study focused on determining the validity and reliability of a hypothesised framework of Arithmetic Word Problem Solving (AWPS) and focused specifically on addition and subtraction problems. In order to establish the validity and reliability of this framework a test was developed and a number of different methods of data collection were used to collect different types of validity evidence.

While addressing research questions, a researcher engages in sampling procedure that involves selecting research site(s), deciding on a participation selection method, deciding how many participants are needed for answering the research

questions and selecting participants who would provide data for the research (Creswell & Clark, 2018).

To gather the necessary evidence for this research, data were collected in the form of students' tests, students' post-test interviews (think-aloud protocols), expert opinionnaires, school questionnaires, and students' test scores from school examinations. There were two phases of the study: the piloting and the final data collection. Wilson (2005) recommends using multiple iterations of an assessment to collect validity evidence that can be used for making the assessment more valid. This can be done through small- or large-scale studies. In this study, the items developed based on the AWPS framework were first administered in a small-scale study, that is, the pilot study (62 student participants) and then the items were revised based on the pilot data and the revised items were administered in a larger scale study (808 student participants), which has been called the final data collection/final study in this thesis.

In order to collect data, first the schools were selected, then the students and finally the experts. This section will discuss the strategies discussed in different research literature for selecting research participants and how the participants of this study were selected.

Before discussing the different ways of selecting participants for a research study, it would be helpful to understand key terms such as population, target population/sample frame and sample. A group of individuals/organisations having the same characteristics are known as population, whereas a group of individuals/organisations having some common defining characteristics that the researcher can identify and study is a target population/sampling frame and finally a

subgroup of the target population that the researcher plans to study is known as a sample (Creswell, 2012). Sampling is the method of selecting a sample. Certain inferences about a population can be made by determining the characteristics of a sample drawn from that population (Best & Kahn, 2014).

The research literature discusses two broad approaches of sampling: probability sampling (also called random sampling) and non-probability sampling (also called non-random sampling).

Researchers select individuals who are representative of a population in probability sampling (Creswell, 2012). These sampling techniques are termed probability sampling because it is possible for the researcher to determine the probability (or chance) of each member of a defined population to be selected in the sample (Mills & Gay, 2016). This in turn helps in using data analysis techniques that can be used to generalise findings from the data to the whole population. Hence, probability sampling is considered to be the most rigorous sampling for quantitative research. However, it should be mentioned that successive random samples taken from the same population are likely to differ in their characteristics but their variation from the population and from each other can be estimated (Best & Kahn, 2014).

There are different types of probability sampling techniques available for education researchers. Most authors (e.g., Best & Kahn, 2014; Creswell, 2012; Creswell & Creswell, 2018; Mills & Gay, 2016) discuss some variations of these four as the four key types of probability sampling: simple random sampling (*selecting participants from a list with the help of a random numbers table*), systematic random sampling (*choosing a random start on a list then selecting every n th person on the list,*

where n equals to the number of people on the list divided by the required sample size), cluster sampling (selecting a cluster e.g., a school or a school district through random sampling from a list; then selecting participants from that cluster through simple random sampling or selecting whole groups or classes if more suitable) and stratified sampling (stratifying the population based on some known characteristics e.g., gender, education; then using simple random sampling to select from the strata). In addition to these, Cohen et al. (2018) discuss stage sampling (an extension of cluster sampling) and multi-phase sampling (similar to stage sampling but the population changes at every phase).

Though probability sampling techniques provide the best methods to select a suitable sample for research, it is not always feasible or possible to use probability sampling in educational research. This is often due to practical constraints or resource limitations. In such cases, researchers often use non-probability sampling techniques. Besides, selection of sample for qualitative data collection does not require a probability sampling and can work with non-probability sampling techniques. Cohen et al. (2018) suggest that, irrespective of quantitative or qualitative approach, research that has limited scope often uses non-probability sampling. Instead of the disadvantages stemming from the non-representativeness, they are less complicated to administer, considerably cheaper, and adequate if the researcher is not aiming to generalise the findings to a wider population.

In non-probability sampling, the researcher selects individuals from the population who are available and are convenient to approach or willingly volunteers and represent some characteristics that she or he intends to study (Creswell, 2012). In

this approach, it is not possible for the researcher to determine the mathematical probability (or chance) of each member of a population to be selected in the sample (Cohen et al., 2018; Mills & Gay, 2016).

Research authors have discussed several non-probability sampling techniques. Creswell (2012) highlights two non-probability sampling techniques as popular among education researchers: convenience sampling (where participants are selected because they are available and willing to participate, and the researcher picks the nearest individual available and continues the process till the required sample size is reached) and snowball sampling (where researcher begins with some participants through convenience sampling and they identify more participants to be part of the sample). Mills and Gay (2016) discuss convenience sampling (also referred to as accidental sampling or haphazard sampling), purposive sampling (also known as judgmental sampling) and quota sampling. Cohen et al. (2018) outline five types of non-probability sampling: convenience sampling, quota sampling (researcher identifies significant characteristics of the population and strives to represent these in proportions similar to the population in the sample), dimensional sampling (researcher identifies different factors of interest in the population and selects at least one participant for every combination of those factors), snowball sampling and purposive sampling (researchers select cases based on their judgment of typicality and possession of specific characteristics).

Based on this review of different sampling techniques, convenience sampling was deemed to be the most suitable for this study. Convenience sampling is a type of non-probability sampling that involves choosing those who happen to be available and

accessible for participating in research. Researchers simply choose the sample from those to whom they have easy access (Best & Kahn, 2014; Mills & Gay, 2016).

Considering the time available for data collection, travel costs and manageability of data, it would not have been possible to use any kind of probability sampling. So, non-probability sampling was preferred. Within the different non-probability sampling techniques, convenience sampling was the most feasible choice because this would provide adequate flexibility to complete data collection within the limited time available. Besides Mills and Gay (2016) suggest that convenience sampling is the most used sampling technique in education research.

Cohen et al. (2018) provide a comprehensive list of factors that should be considered when determining the sample size for a research study. Based on their list, a number of factors were considered while deciding the sample size for each of the instruments in this research. These include research purposes, question and design, the size and nature of the population, consideration for attrition and non-response.

3.4.1 Population

The target population for this research is Bangladeshi students studying in grades 4 and 5 (age 10 – 11). The Bangladeshi primary level was selected because of the researcher's familiarity with the context and his passion for contributing to his country. As per the Bangladeshi primary curriculum, students are taught arithmetic word problem solving in grade three (age 9). In order to collect data from students who should have at least some experience of doing arithmetic word problem solving, students of grades 4 and 5 were targeted for this research. Furthermore, to make it

manageable for the researcher data collection was limited within the Narayanganj district of Bangladesh.

3.4.2 Selection of sample for the pilot study

As mentioned earlier, the field work of this study had two phases: piloting and final data collection. In the piloting phase, the student test (with pilot items) was administered, data about the school were collected through the school questionnaire, data on students' performance on school tests in Bangla and Mathematics were collected and students' cognitive interviews were conducted.

The researcher contacted seven schools in the Narayanganj district of Bangladesh at the beginning of 2017 and invited them to participate in the study. These schools were selected based on convenience sampling as they were all located within an easy travelling distance for the researcher. By June of 2017 three schools had communicated their interest to participate in the study. As the piloting was planned in July 2017, one of these three schools was selected for piloting of the items.

This school has a set of attributes that made it an ideal candidate for piloting. First, it is located in a semi-urban area where students from different socio-economic statuses attend; second, it is a co-education school where students of both genders were available for the test; third, as per the Primary Education Completion Examination (PECE) results for the targeted region, this school presents a good mix of different ability students, and finally, the teachers and administrators of this school were very co-operative and enthusiastic about participating in the study. All the students studying in grades 4 and 5 of this school were selected as the sample for administering the pilot test. The school had 35 students in grade 4 and 30 students in

grade 5. So, the total sample would have been 65 students. However, on the day of testing, 33 students were present in grade 4 and 29 in grade 5. All of those who were present participated in the test so the total number of students taking the test was 62. This constituted the pilot sample for the quantitative data from the test. In addition, students' scores on Bangla and Mathematics tests in the school examinations were collected and the school questionnaire was administered.

According to the demographic information collected from the participants, 42 (67.7%) were girls and 20 (32.3%) were boys. About 80% of the students mentioned they have attended pre-primary school and 78% said they took tuition outside of school. Both grades had a male mathematics teacher who had 15 years of education (a bachelor's degree) without any mathematics or science specialisation and had 18 years of teaching experience. The participating students had a varied socio-economic backgrounds, with their family's average monthly expenditure ranging from 2000 to 100000 taka (approximately 34 AUD to 1666 AUD).

Qualitative data were collected through post-test cognitive interviews. Johnstone et al. (2006) argue that data collection in such interviews is labour intensive and the sample size is usually small. Since the research process is intensive, even small samples can provide valid information. Previous researchers (e.g., Nielsen, 1994) have suggested that even five participants may be able to provide adequate information about problem solving behaviour. In the pilot phase of this research, six students were selected from the same school through convenience sampling to conduct cognitive interviews. Three of them were from grade 4 and three from grade 5. The researcher tried to get students of varying ability for the interviews so perspectives of different

ability students could be gathered. In addition, the researcher ensured inclusion of both genders in the sample for the interviews.

3.4.3 Selection of sample for the final data collection

The final data collection included quantitative and qualitative measures. These will next be discussed separately.

3.4.3.1 Sample selection for quantitative measures

There were a number of quantitative measures on which data were collected in the final data collection of this study. The most prominent data collection instrument was the test of students. This test included items from the five strands of the arithmetic word problem solving framework and items from textbooks. In addition, school questionnaires were used to collect background data on the schools. Furthermore, data for students' school tests in Bangla and Mathematics were collected. Apart from these, the other quantitative measure was the expert opinionnaire. This section first discusses the sample selection for the measures based on schools and then discusses selection of a sample for the expert opinionnaire.

As mentioned earlier, the researcher contacted seven schools in the Narayanganj district of Bangladesh at the beginning of 2017 and invited them to participate in the study. These schools were selected based on convenience sampling. By June of 2017 three schools had communicated their interest to participate. The piloting of the items was conducted in one of these schools in July 2017. The final data collection was conducted from August to September 2017. Some of the schools that were contacted initially to participate in the final phase of data collection were

reluctant in responding to the invitation. Due to time constraints three additional schools were contacted during July-August of 2017 and they agreed to participate. Once the data collection started in the schools that had already agreed to participate there was some communication among the head-teachers of the schools (as they are all under the same school administrative region), and then the schools that did not respond to the invitation initially also wanted to participate. As a result, nine schools were included in the final data collection for this study. All these schools are in the Narayanganj district of Bangladesh. All the students studying in grade 4 and grade five of the participating schools were the target sample of the study. Some schools have more than one section in each grade. In some schools, data were collected from all sections in a merged class while in other schools, data were collected from a single section. The reason behind this will be further explained in the administration section.

Although the schools were selected through a convenience sampling, the researcher tried to ensure a good mix of typical background variables such as school types and geographical locations. Among the nine schools in the final phase of data collection seven were Government primary schools (which is the prevalent type of primary school in Bangladesh), one was a privately managed school and the remaining one was a primary school attached to a high school. In terms of the schools' self-reported geographical locations, two were from rural areas, four from semi-urban areas, and three from urban areas. However, it should be kept in mind that the schools were from a single district so the varying geographical locations might not represent an actual nationwide variation. To get an understanding of the socio-economic status of the respondents, data were collected on annual school fees families have to pay. This ranged from 0 to 2800 taka (\$47 AUD) for grade 4 and from

0 to 3200 taka (\$54 AUD) for grade 5. Students in the government primary schools do not have to pay any annual fees, so for 79% of the respondents the annual school fee was 0 taka. Mathematics class contact time per-week for the schools ranged from 210 minutes to 270 minutes.

Table 3.3

Number of Respondents from Each School for AWPS Test in the Final Data Collection

School Research ID	School Type	School Location	Grade 4	Grade 5	Total
02	Government	Semi-Urban	26	36	62
03	Private	Semi-Urban	74	52	126
04	Government	Semi-Urban	41	48	89
05	Government	Urban	37	27	64
06	Government	Urban	61	53	114
07	Government	Urban	65	81	146
08	Government	Rural	45	45	90
09	Government	Rural	43	46	89
10	High school-attached	Semi-Urban	30	16	46
Total			422	404	826

Note. School research ID 01 was assigned to the school where data collection for the pilot study took place.

Data were collected from both grades 4 and 5 in all the participating schools. Table 3.3 gives an overview of how many students took the test in each grade in each school. In total, 826 respondents took the test in the nine participating schools. They represent

the sample of data collected for the AWPS test items, the textbook items integrated within the test and student score in school administered Bangla and Mathematics tests.

As the items were based on the Bangladeshi context and curriculum and were administered in Bangladesh, it was felt that opinion of Bangladeshi experts regarding the items would be most suitable. In Bangladesh most tests are developed by schoolteachers themselves and there are no subject specialist teachers at the primary level. Any teacher can be given the responsibility of teaching any subject. However, there are some experts who work with mathematics as a subject in the apex academic body of primary education – the National Academy of Primary Education (NAPE). In addition, there are some instructors in the Primary Teachers Training Institutes (PTIs) who train primary teachers in Mathematics Education. For this research, NAPE specialists, PTI instructors, and primary teachers (both head teachers and teachers) were considered as experts and expert opinion data were collected from them. Data were collected from 24 experts in total. Most of them were primary teachers or head teachers. All respondents were selected through convenience sampling. The distribution of different types of experts is shown in Table 3.4:

Table 3.4

Respondent Sample Distribution for Expert Opinionnaire

Respondent type	N
Specialists from National Academy of Primary Education (NAPE)	2
Primary teachers' Training Institute (PTI) Instructor	6
Primary school head teacher	3
Primary teacher	13
Total	24

Analysis of the demographic variables suggests that there is a wide variety of educational backgrounds among the experts, with a minimum of Higher Secondary Certificate (also known as HSC, which entails 12 years of schooling) to a maximum of Masters/MSc/MPhil. Except for the one respondent with HSC qualifications, all the others had at least a Bachelor's degree. More than 50% of the respondents (14) had at least a Master's degree and 50% (12) had some form of teaching qualification. Their teaching experience also varied widely with a minimum of one-year experience and one respondent suggesting s/he had 50 years of teaching experience. The respondents' mean experience was 11.07 years and the median experience was 6.00 years.

3.4.3.2 Sample selection for qualitative measures

The only qualitative instrument that was used for collecting data was the cognitive interview. Mills and Gay (2016) suggest that qualitative sampling entails selection of a small number of participants for a study who are good key informants and are able to improve the researcher's understanding about the research topic. As mentioned earlier, because of the labour-intensive nature of data collection through cognitive interviews, the sample size is usually quite small (Johnstone et al., 2006). Some experts (e.g., Nielsen, 1994) have considered five participants to be adequate where others (e.g., Sheatsley, 1983) have suggested 12-25 participants are enough for revealing difficulties and weaknesses in a questionnaire. Beatty (2004), for example, selected 18 participants as sample for cognitive interviews. Considering these

examples, it was decided that cognitive interviews would be conducted with 20 students for this research.

Qualitative data were collected from three schools in the final phase of the study. Both the selection of schools and the selection of students were done through convenience sampling. However, while selecting the students, several factors were considered. Students' ability level based on their merit position in the class was the major factor, as the intention was to have interviews with students from a range of ability groups. Another factor was the respondents' gender. Effort was given to include equal numbers of boys and girls. Although one of the schools where the interviews were conducted was a boys' school, more girls were selected from another sample school to offset this. These factors were considered to avoid bias in the interview sample (e.g., having more high-ability students might make it difficult to understand some students' difficulty in answering some items, or including a disproportionately large number of boys/girls in the sample might introduce a gender bias in the results). Table 3.5 provides an overview of how the participants of the interviews were distributed.

Table 3.5

Distribution of Students Selected for the Cognitive Interviews across Different Schools in the Final Phase Data Collection

School RID	Grade		Gender		Total
	Grade 4	Grade 5	Female	Male	
3	4	4	4	4	8
2	1	3	0	4	4
9	4	4	7	1	8
Total	9	11	11	9	20

In total 20 interviews were conducted in three schools. Among the respondents, nine were grade 4 students and 11 were grade 5 students. Eleven of the respondents were girls and nine were boys. The data collection was limited within the three schools because of time constraints (i.e., more time would have been needed to do the 20 interviews in a larger number of schools).

3.5 Administering the data collection instruments

3.5.1 Data collection for pilot

The school where piloting was conducted was one of the first schools to respond to the researcher's request to be part of the study. Once formal permission was secured from the school's authority, the researcher had a discussion with the head teacher and other teachers in the school explaining what the study was about and what the schools' participation would entail. In addition, the school questionnaire was given to the head teacher to collect background information about the school.

To recruit grade 5 and five students for the study, the plain language statement (PLS) and parents' consent form for their children's participation in the study (both written in Bangla) were distributed in grades 4 and 5 with help of the class teachers. The consent form sought the parents' permission to recruit their children for both the test and the post-test cognitive interviews because at that point it was not determined which students were going to be interviewed. In addition, permission was requested to collect students' mathematics and Bangla score from the school tests. Students were requested to take the PLS and the consent form to their parents and

get their permission to participate in the test. In case the parents had any questions regarding the study or the data collection, they were advised to contact the researcher or the school's head teacher to get more information. The students were asked to bring back the signed consent forms by the following week. The English PLS and consent form is included in Appendix 3.15 and the Bangla PLS and consent form (which was used) is included in Appendix 3.16. In addition to providing the details to the parents, students were given a plain language statement (PLS) which briefly explained the nature of their possible participation in the research. The PLS explained what the test and the interview were going to be like and how the results were not going to affect them in any negative way (English translation included in Appendix 3.17 and Bangla version in Appendix 3.18).

Based on the schools' suitability of schedule, a date was fixed during the following week for administering the test. A test booklet was prepared with the final set of pilot items and an adequate number of copies were printed taking into consideration the number of students in the two grades and adding 10% extra booklets in case there were errors in some of the booklets.

All students returned with their signed consent forms and on the day of the testing all those who were present participated in the test. They were also given an assent form (English translation in Appendix 3.19, Bangla version in Appendix 3.20) before starting the data collection. They were given a choice of taking the test or moving to a separate room where they could do other activities (e.g., doing some supervised reading) and told that they could stop taking the test at any point if they felt uncomfortable. However, all of them were enthusiastic about participating in the

test and no one wanted to miss out. In addition, before starting the test, students were told that a few of them would later be interviewed about how they answered the test items. They were asked to show by raising their hands if they would like to participate in the interviews. All of them were keen to participate and raised their hands.

The students were seated in their own classrooms in a test-taking set up. Both grades took the test simultaneously. As this was the piloting, students were given the liberty to take their time so that the researcher could determine how long it usually takes for them to finish the test. The students were told that they had approximately an hour, but they should feel free to take more time if needed. It took them 50 to 70 minutes to finish the test. Most of them finished within 60 minutes and only three students needed more time. All of them were completed by 70 minutes. The researcher observed both rooms (which were located side-by-side) to see if students were comfortably answering the items. Both rooms were attended by a teacher from the school to make sure the students were answering the items independently. At the end of the test, all the test booklets were collected by the researcher. In addition, students' scores on school administered Mathematics and Bangla tests were collected from the head teacher.

Post-test cognitive interviews were conducted with six students to collect qualitative data. All the interviews took place within five working days of the test administration and were conducted by the researcher. The school has a room for pre-school children that was available during the second shift of school hours when grades 3 to 5 attend. This room was used for doing the interviews. The participants of the

interview were selected with help of the class teacher and interviewed individually during class breaks. The class teacher introduced the researcher to them, and the researcher asked them if they would like to be a part of the interview. After getting their assent, they were invited to the interview room and the interview was conducted. They were given the pilot test and requested to articulate their thinking while solving the items. After finishing the items, they were asked some additional questions (e.g., which item within the test they found to be difficult to answer? If they faced any trouble in understanding any item? etc.). The English translation of the fieldnote template used for the cognitive interviews in the pilot study is included in Appendix 3.21. The Bangla version (which was used) is in Appendix 3.22. After completing the interview, they were thanked for their time and taken back to their classroom.

The researcher took fieldnotes while the students were thinking aloud. The interviews were conducted in Bangla. The notes and some verbatim quotes that were written down were later translated to English. Each interview took approximately an hour to conduct, and after the interviews the researcher took some time to make sure all the field notes were jotted down properly. Simple remarks or codes that were written by the researcher during the interview were elaborated right afterwards to ensure validity of the collected information.

3.5.2 Final data collection

This section discusses the activities in the final phase of data collection for the study. The discussion is divided into two broad sections: collecting quantitative data and collecting qualitative data.

3.5.2.1 Collecting quantitative data

The final data collection took place after the instruments were revised based on the piloting. In total, nine schools participated in this phase of data collection. After securing formal permission from the participating schools, the researcher had discussions with the head teacher and other teachers at the participating schools explaining what the study was about and what the schools' participation would entail. The school questionnaires were given to the head teachers to collect background information about the school and teachers teaching Mathematics in Grades 4 and 5. The completed questionnaires were collected from the schools in later visits.

All students studying in grade 4 and grade 5 of these schools were in the sample for data collection. In large schools, where there is more than one section available for each grade, the head teachers' advice was sought regarding the feasibility of testing all the students. The data collection took place during the monsoon season and many students were skipping school due to the bad weather. In such cases, the schools were merging sections to a single class and the head teachers advised to approach parents of students in all sections for consent. Once consents were secured, it would be possible to administer the test on the scheduled day with a single section if the sections were separated or with all the students if the sections were merged into one class. So, in some schools, data were collected from all sections in a merged class while in other schools, data were collected from a single section.

To recruit the students as respondents, a plain language statement (PLS) and parent's consent form for their children's participation in the study (both written in Bangla) were distributed in grades 4 and 5 with help of the class teachers (English

translation in Appendix 3.15 and Bangla version in Appendix 3.16). Similar to the piloting, the consent form sought the parents' permission to recruit their children for both the test and the post-test interviews. In addition, their consent was sought to collect students' test scores in Mathematics and Bangla (language) in the school tests. The students were advised to contact the researcher or the school's head teacher to get more information if their parents had any questions regarding the study or the data collection. Students were also given a plain language statement about the research (English translation included in Appendix 3.17, Bangla version in Appendix 3.18) briefly explaining what the research was about and how participating in the test or the interview were not going to affect them in any negative way.

After sending out the PLS and consent forms, a date was fixed within the next few days depending on the suitability of schedule for the schools. A test booklet with the revised test items was prepared and an adequate number of copies were printed considering the number of students in the target grades plus 10% extra for any mishaps (e.g., printing or binding error in the test booklets). The signed consent forms were collected by the head teachers prior to the testing dates. On the day of testing, students were given an assent form (Appendix 3.20, English translation included in Appendix 3.19). They were given a choice of taking the test or participating in an alternative activity (e.g., supervised reading in another room) and told that they could stop taking the test at any point if they felt uncomfortable. The students were seated in their own classrooms in a test-taking set-up. They were given 60 minutes to answer the 38 items in the test booklet. The test took place during their mathematics classes, and the mathematics teacher helped with the invigilation to make sure the respondents were answering the items independently. In each school the test was

administered in both grades on the same day. Due to scheduling conflicts, it was not possible for the researcher to be present during the test administration in two schools. A research associate who was familiar with this research project from the design phase helped out with the test administration in those schools. This associate has a postgraduate degree in Educational Evaluation and Research and currently works as a trainer for teachers of primary level.

While visiting the schools for administering the tests, students' scores in Mathematics and Bangla (language) in school examinations were collected. A pre-printed template was given to the class teachers where students who appeared in the AWPS test were identified. Teachers were requested to provide those students' Mathematics and Bangla scores in the latest school examination.

As mentioned earlier, an expert opinionnaire was used to collect data about the items' validity. This took the most time to administer as the participants were from many different organisations. After identifying potential respondents, the researcher approached them regarding data collection and sent them a PLS and consent form (Appendix 3.25). Some of them were contacted face-to-face while others were contacted via emails and sent the documents online. Once they had agreed to participate in the study, they sent their signed consent form to the researcher and completed the opinion survey. Some of them were prompt in replying while others took more time. In total, 24 completed expert opinionnaires were collected by the researcher by the end of 2018.

3.5.2.2 Collecting qualitative data

Qualitative data were collected about the AWPS test items in the form of cognitive interviews. While administering the test, students were asked if some of them would like to participate in an interview about the test they had taken. All of them showed interest. After the test was finished, a convenience sample of students were chosen with help of the schools' mathematics teachers to participate in the interview. Twenty interviews were conducted across three schools. Among these, 16 were conducted by the researcher and four were conducted by the research associate mentioned earlier. She was familiar with the research from the design phase and helped with the quantitative data collection too. Before she conducted the interviews, the researcher arranged a training session with her where the details about the research and the cognitive interviews were discussed thoroughly. She also conducted a mock cognitive interview as part of this training session.

A standard operating procedure was developed before conducting the interviews to make sure all the interviews across the different schools are conducted in a similar fashion. It was thought that this would be helpful in gathering valid and reliable data. This very simple procedure is briefly discussed here.

Most of the time, the interview took place during class breaks of the students. As mentioned earlier, parents' consent was collected for all the students participating in the quantitative data collection (PLS and consent form included in Appendix 3.15 [English] and Appendix 3.16 [Bangla]). The same consent letter had a provision for including their children in an interview. All children participating in the quantitative data collection had parents' consent for qualitative data collection too. While conducting the interviews, first, the selected student was asked if s/he would like to

participate in the interview. If s/he agreed (most students agreed), s/he was brought to the room where the interviews were conducted. Almost all primary schools in Bangladesh have a separate room for pre-primary students which usually remains unoccupied during the time classes for grades 3 to 5 take place. With permission of the school authority, this room was used for the interviews. The student was seated comfortably, and the interviewer introduced himself to the student and explained the purpose of the interview and how it was not going to affect the student in any adverse way. The 'warm-up – interview – cool-off' routine suggested by Dowling and Brown (2010) was followed for the interviews where the conversation began with exchanging pleasantries (e.g., asking 'how are you?'), then moved on to discussion about the items and finally was wrapped up by summarising the last answer of the participant and thanking her/him for participating in the study. They were also asked if they had any questions about the interview or the research before finishing. Furthermore, if the student seemed nervous, s/he was asked some unrelated questions to make her/him more comfortable. Such questions included: what their favourite subject was; what they liked about those subjects; what their favourite pastime activity (hobby) was; what sport they liked, and so on. These familiar topics seemed to ease them and warm them up to the interviewer. After they looked comfortable enough, the interviewer explained how the think-aloud interview was going to take place. The interviewer took the example item included in the questionnaire and demonstrated how to think aloud while solving that item. Then the student was asked to similarly think aloud for the items s/he had already answered while seating for the test. If s/he still had trouble understanding what s/he is supposed to do, the interviewer explained

the think-aloud process again with another example (one of the textbook items from the test booklet).

The student would then read the items herself/himself, describe their thought process while solving each item and finally select an answer. The interviewer asked the student probing questions if there was any confusion about how the student reached the answer, that is, if clarification was required about the thinking process. It was envisioned that for some of the more difficult items, some students might not be able to reach the answer or articulate their thinking processes. It was therefore decided that when a respondent seemed to be struggling with an item, s/he would be asked probing questions and if it was not possible for them to answer the item and if they remain silent for 15 seconds, the interviewer would request them to move on to the next item. This procedure has been suggested by Gadermann et al. (2011). After going through all the items, the interviewer thanked the respondent for their time and took them back to their class.

The interviewer took note of the respondents' responses against the items in a fieldnote template that had all the interview items with spaces allocated under them for writing notes (English translation included in Appendix 3.23, Bangla version in 3.24). These notes also included some verbatim quotes as deemed important. After finishing the interview, the interviewer went through the notes and immediately wrote details where small notes were taken. This was done right after the interview to ensure the notes were made with a fresh memory. In addition, some summary comments were also written on the same day as the interview.

All the interview notes were taken in Bangla. These were then translated into English for the analysis, verbatim quotes were also translated. A sample of the translations were cross-checked to ensure translation validity. All translations and cross-checks were done by the researcher.

3.6 Data management

This section discusses how the data collected through different instruments in piloting and final data collection were managed. This includes description of the different databases that were developed, how quantitative data were prepared and organised, what score templates were used and how error analysis was conducted for data cleaning before going into analysis.

3.6.1 Data management for pilot

Once the test booklet was finalised for the piloting, a template was prepared in SPSS for data entry. This template had the student background variables as well as the item responses. After administering the test, the test booklets were collected by the researcher and the open responses were scored. All correct responses for the open-ended items (3, 9, 15, 21, 27, and 33) were coded as A. Student responses were then entered in the SPSS template by the researcher. A 10% error check was conducted to make sure the data entered in SPSS matched students' responses. The finalised SPSS data file was used for running descriptive analysis in SPSS and for doing item analysis in *ConQuest*.

Students' interview notes were written in Bangla during the interview, which were then translated to English. The verbatim quotes that were collected were also translated.

3.6.2 Data management for final data collection

Quantitative data from schools

This set of data from schools included students' responses in the test booklet, their scores from school administered examinations and the schools' background information. Before collecting data, each participating school was assigned a two-digit numeric Research ID (from 02 to 10) (mentioned in Table 3.3) for ease of organising the data (the school where piloting took place was assigned numeric ID 01). The numeric school Research IDs were printed on the test booklets before distributing them to the respondents. This ensured test booklets from different schools did not get mixed-up due to clerical error.

After administering the tests, the first step was to check the test booklets. This was done in order to ensure that the respondents had put down the necessary basic details such as their grades, sections, roll numbers, and so on. Without these details it would not be possible to make sure their consent forms were in order and it would not be possible to compare their test results with those of the school tests. Once these details were checked, the next step was to assign a Student Research ID to each respondent. This ID contained 6 digits, where the first two digits were the respondent's school's research ID (02 to 10), the third digit was the respondent's grade (4 or 5), the fourth digit was the respondent's section in that grade (0 if there was no section), and the final two digits were the respondent's roll number in the class. So, the six-digit ID WWXYZZ consisted of School Research ID(WW)-Grade(X)-Section(Y)-Student Roll Number (ZZ). Assigning these IDs made it possible to maintain

the privacy of the respondents, easily compare test scores to school-based examination scores, and identify specific cases within the dataset when necessary.

The next step was preparing the data for entry into a database. This included setting up some ground-rules for data entry, assigning codes to responses, preparing a template for data entry into SPSS and preparing a codebook. Brown and Farruggia (2012) advise creating a codebook before starting the data entry activities and emphasise the importance of a codebook in ensuring correct values are being assigned to responses.

A sample of the test booklets was reviewed to set up the ground rules for data entry. Based on this review the following ground rules were decided:

a) Some students ticked or marked an answer on the right side of the options, or just ticked on the word or option itself instead of ticking on the letters as instructed. Since they were perfectly understandable, they would be accepted as valid answers.

b) In cases where students selected one answer, crossed it out and selected another, the answer would still be accepted. In some cases, students had marked and crossed out multiple answers. Even in such cases, the finally selected answers would be accepted.

c) In cases where students selected multiple options, did not cross-out any and it was not possible to determine which of the options they intended to present as the final answer, the item would be marked as 'missing' (99). One exception to this was Item 13 where there were two correct answers, so if a respondent has selected both the correct answers, one of the

letters would be entered so the item would be recorded as “correct” in the analysis.

d) In cases where it was clear that the respondent misplaced their parents’ profession or educational qualification, the placements would be corrected during data entry. For example, in some cases, students wrote their parents’ profession in the space provided for parents’ education, and vice versa. These would be entered in the data base correctly.

e) In some cases, the respondents misspelled their parents’ educational degree/qualification or their profession. As long as the qualification or profession was understandable beyond any doubt, they would be corrected and then entered. However, if the misspelling became so unintelligible that it didn’t make any sense, the variable would be left empty.

f) In many cases, while noting their year of birth, respondents missed one digit. For example, instead of writing 2006 they wrote 206; or instead of 2005, they wrote 205. When teachers were consulted regarding this, they suggested that students sometimes do this when writing down years. So, these would be corrected to the appropriate number of digits and entered.

Table 3.6

Codes for Open Ended Items (03, 09, 15, 21, 27, 33)

Response type	Assigned code
Correct use of strategy (i.e., the correct answer)	A
Incorrect strategy use or mistake in calculation	B

Using wrong operation (e.g., addition instead of subtraction)	C
Putting down wrong numbers	D
Response missing	99

Reviewing a sample of test booklets also helped in assigning codes to responses. For some of the demographic variables, such as school research ID, student research ID, student grade, student birth year and so on, no codes were assigned. It was decided that it would be easier to enter the information as it is. For other variables such as student's section, student's gender, if the student has attended pre-primary education, and so on, codes were assigned. Details of these codes can be found in the code book (Appendix 3.26).

Most of the test items were multiple-choice questions. Hence, the responses were limited to A, B, C, D, or E and they were entered as letters. In addition, '99' was assigned as the code for missing responses. Some items in the test were open-ended (i.e., items 3, 9, 15, 21, 27, and 33). Items that required judgment for determining score needed specification of a scoring rubric (Wolfe & Smith, 2007). For these items, a set of codes were decided before the data entry began. The booklets were checked and based on the respondent's answer, codes were assigned on the items in the booklet. These codes are described in Table 3.6.

After codes were decided for different responses a template was prepared in SPSS with all the 38 items (32 items for AWPS+6 textbook items) and the other variables. In total, the initial template had 52 variables which had a mix of numeric and string entry types in nominal and ratio measurement scales. Details about the template can be found in the codebook (Appendix 3.26).

Another level of coding took place after the data were cleaned, entered in SPSS, and checked for errors. Based on the answer key for the items, the letter responses were recoded to 1 for the correct answers and 0 for the incorrect ones. This was done with help of the recoding function in SPSS.

In addition to the AWPS test, two more datasets were collected from the schools: school background information (e.g., geographical location, school type, etc.) and students' Mathematics and Bangla test scores. Two separate templates were developed for these data in SPSS and data was entered. Schools' information was entered along with the school research IDs assigned earlier and students' Mathematics and Bangla test scores were entered along with their research IDs. As a result, it was possible to later connect the school information and student scores from school tests with the AWPS test scores.

3.6.3 Date entry and data cleaning

3.6.3.1 Students' test data

Once the codes and the template were finalised the data entry began. All the data were entered by the researcher himself. There were 826 respondents in total for the test. During the process of data entry 18 test booklets were found to be incomplete. They were missing crucial details such as student's roll number or were missing most of the item responses. These incomplete booklets were not entered in the database. As a result, the total number of cases in the database became 808.

After entering data, it is essential that researchers check the data for errors (Creswell, 2012) and perform data cleaning. Data cleaning means checking the data

for anomalies and rectifying those (Wu et al., 2016). It ensures that human transcription of paper-based responses to digital data has been done accurately and, in addition to raising confidence that accurate representations of participant responses are analysed, gives an opportunity for more robust inferences and conclusions (Brown & Farruggia, 2012). Two methods of data cleaning were employed for the test data. First, a frequency distribution was generated through SPSS to check all the values were within the acceptable ranges (as discussed by Creswell, 2012 ; Wu et al., 2016) and then to further assess the data for human error during data entry, an error check protocol was administered with the database. Using SPSS, a 10% sample of the entered data was selected randomly. The software gave a sample of 83 cases. These cases were checked against the booklets for error. In total, 4316 entries were checked (52 variables per case × 83 cases) and only 9 errors were found. The percentage of error in the entries was calculated to be 0.20%. The 9 errors are described in Table 3.7. These errors were amended in the main dataset.

Table 3.7

Data Entry Error Check Results

Respondent ID	Error
RID 25030	Item number 26, entered as B. The tick seems to be in between A and B; probably more towards A. Understandable how it might have been confused as B.
RID 25030	Item number 31, entered as E. The tick seems to be in between D and E; probably more towards D. Understandable how it might have been confused as B.
RID 25032	Student pre-primary attendance marked as 'no', where it should be 'yes', student private tuition marked as 'yes', where it should be 'no'
RID 35062	Item number 38, Student selected C, then crossed it out and selected B, but data was entered as A.
RID 44038	Item number 10, Student selected A, data entered as B.

RID 45003	Student entered '1' in Private tuition variable, entered as blank.
RID 54129	Item 17, Student marked 'C', Entered as 'E' in Database.
RID 54218	Item 22, Student marked 'E', Entered as 'C' in Database.
RID 64276	Item 36, Student marked 'C', Entered as 'B' in the Database.

3.6.3.2 Quantitative data from the expert opinionnaire

Based on the variables included in the expert opinionnaire a template was prepared in SPSS for entering the data. After completion of data collection, all the information was entered in SPSS. A frequency distribution check for unlikely values and a 10% error check for ensuring the opinionnaires were entered accurately were done.

3.6.3.3 Qualitative data from students' cognitive interviews

After completing all 20 interviews, the responses were collated. Separate files were created for each item, where all the responses from the 20 students for an item were brought together. Information on respondents' grade, their merit position in their classes and their gender were retained in this collation through codes.

3.7 Data analysis

This section discusses how the data collected through different instruments in piloting and the final data collection were analysed. Similar data analysis techniques were applied to evaluate the psychometric properties of the test administered in the piloting and the final data collection. However, their results are mentioned later in separate chapters. In addition to discussing the item analysis for the piloting and final test, this section discusses the analysis of other quantitative instruments used in the

final data collection (i.e., expert opinionnaires, students' scores from the textbook AWPS items in the test, and students' scores in their school mathematics and Bangla tests). Furthermore, analysis of qualitative data from students' cognitive interviews is also discussed.

3.7.1 Psychometric analysis of the test in piloting and final data collection

In order to make meaningful inferences from the data collected by the test, the third and fourth building blocks of Wilson's (2005) model were utilised. The third building block is the outcome space (gathering and scoring item responses) and the fourth building block is the measurement model (producing measures).

The first step in preparation of data analysis was defining the outcome space for the test items. The outcome space for a test is described as the decisions about aspects of the response to items that will be used as the basis of inference and how they are categorised and scored (Wilson, 2005). This means representing qualitatively different response categories in numeric codes which can be dichotomous (two qualitatively separate categories/levels) or polytomous (more than two different categories/levels) (Wolfe & Smith, 2007).

3.7.1.1 Gathering and scoring item responses

For this research a dichotomously scored outcome space was used. All items in the test were dichotomous, so the outcome space consisted of 0s and 1s where 0 represents an incorrect answer and 1 represents a correct answer. It should be mentioned that such scoring represents students' location in an ordinal scale of measurement where the higher number means more of the construct is being measured. The scoring was done with the recoding function in SPSS. However, the

data file used in *ConQuest* included the unscored categorical responses, so the answer key was used in running *ConQuest* analyses.

3.7.1.2 Producing measures

The first step of data analysis was producing measures from the test responses. This constitutes the fourth building block from Wilson's (2005) model and is also known as measurement model or psychometric model. A measurement model helps to relate the scored outcomes from the items to the construct(s) the items represent (Wilson, 2005). There are theoretical differences between classical test theory (CTT) and item response theory (IRT) in making sense of data collected about the level of a measured construct. In CTT approach a person's true score (which theoretically can be gathered by multiple parallel tests) shows the level of construct achieved whereas in IRT approach, a latent trait (i.e., the construct) is defined through the model and an individual's scores and item responses reflects his/her position or level in the construct (Wu et al., 2016). Furthermore, in the IRT approach ordinal data codes from scoring are processed with help of a measurement model to produce a continuous scale that can represent achievements of different quantities of the target construct on an interval scale. There are many IRT approaches that are used for various types of tests. For the test used in this research, Rasch's dichotomous model, which is regarded as a one parameter IRT model, was deemed appropriate. The Rasch model is considered as the simplest IRT model and hence, most strict, as the item response function is determined by the item difficulty parameter only (Wu et al., 2016). This model is suitable for tests with items that have only two levels of performance (Wolfe & Smith, 2007).

The Wright map

The construct map discussed previously can be combined with Rasch model to produce a Wright map. Wright maps help visualise item location and respondent location in the same graphical representation and can be used to describe the probability of item responses for students. Both item difficulty and person ability can be depicted in the same scale. The respondents are depicted on the left side and their responses on the right side. The measurement unit of the scale for estimating these two characteristics is known as logit, which is short for “log of odds unit” (Wu et al., 2016, p. 110). The respondent location is generally called the respondent ability and item location is generally called item difficulty, where the difference between these two regulates the probability of the respondent making that response (Wilson, 2005). When the respondent location is above the item difficulty, the difference is positive and there is more than 50% probability that the respondent will make that response. When the respondent location is below the item difficulty, the difference is negative, and there is less than 50% probability that the respondent will make that response. In addition to giving the opportunity for comparing respondents with their responses, the Wright map enables us to compare the responses to different items, since all the items in the instrument are available in the same graphical representation. The relative locations of the items in the Wright map can be compared to the construct map to check how the difficulty assumptions were reflected in empirical data, which in turn may provide construct validity evidence (Wilson, 2005).

Item fit

In the IRT approach, a test is developed based on a defined latent trait or construct where the items in the test are tapping into manifestations of levels of the construct achieved by the respondent. There could be many items that tap into the same construct. In order to make sure that the items that have been chosen for a particular test actually measure the construct we are interested in, we have to measure how the items fit the underlying model. There is an underlying mathematical model in IRT to predict the likelihood of item responses and statistical tests of fit can be used to determine whether the responses of an item fit the IRT model (Wu et al., 2016). Wilson (2005) highlights the technical expectations built into the analysis of instrument data when using a Rasch model, which ought to be checked to make sure that the estimates used are appropriate for the purpose.

There are different types of fit statistics. Among them, residual-based fit statistics were used in this research. Residual-based fit statistics may be weighted (infit mean square) or unweighted (outfit mean square). The weighted mean-square or the infit data provides valuable information regarding how individual items in a test fit the Rasch model. The fit mean-square value has an expectation of 1. There are two possibilities when an item shows misfit: if the mean-square value is less than one, it is considered as an over-fit, and when mean-square value is greater than one, it is considered as an under-fit. The over-fitting items are relatively highly discriminating whereas the under-fitting items are relatively less discriminating (Wu et al., 2016). In addition to infit mean squares, another item statistic called weighted-t statistic (which uses a transformation to convert the weighted mean square into a standard normal

distribution) is often used. It is recommended that both infit mean square and t statistics be used simultaneously to identify problematic items (Wilson, 2005). Infit mean squares and t-statistics calculated with *ConQuest* were used in the analysis of test items in this research.

Respondent fit

Similar to checking whether the items fit the expectations of Rasch model, it is important to identify if the responses to the items meet the expectations of the model. When administering a test that has lots of multiple-choice items, it is easy for students to get distracted or to randomly select an answer for some of the items. Weighted Likelihood Estimates (WLE) help us to identify respondents who had unusual response patterns in a test. Note that , in the context of respondent fit, “overfit” is not usually considered a problem (Wilson, 2005). *ConQuest* was used to calculate the respondent fit for this research.

Item total correlation

Item rest correlation gives us an opportunity to explore how individual items have contributed to the overall test. Items with positive item-rest correlations are suitable for a test, whereas low or negative item-rest correlation means the items need to be reviewed and revised. Item rest correlations were calculated for all items using *ConQuest*.

Transformation of logit scores

Another analysis from the test data was the students’ scores in the AWPS test determined in logit scale using *ConQuest*. However, logit values can be zero or

negative and can be difficult to interpret for some users of assessment results. For instance, “zero logit does not mean there is no ability. A negative logit value does not mean worse than no ability” (Wu et al., 2016, p. 125). To avoid such confusions with logit values, these values are often converted using a linear transformation. The resultant scores (often called scaled scores) are “typically positive and above 100 so they are less likely to be confused with percentages of items correct.” (Wu et al., 2016, p. 125). For the student ability estimates (i.e., logit values) calculated through *ConQuest*, the following linear transformation was used: Scaled Score = Ability estimate X 100 + 500

As a result of this conversion, scaled score for all students fell within a range of 200 to 800, eliminating any possible confusion with zero and negative logit values. And since this is a linear transformation, the interval property of the scores is preserved (Wu et al., 2016).

3.7.1.3 Reliability

In this research, one set of items were administered only once (pilot items were administered in the pilot study and the revised set of items were administered once in the final study). As discussed in Chapter 2, in such circumstances (where one set of items are administered only once), a measure of internal consistency is the most suitable way of determining reliability. Therefore, a number of internal consistency measures were calculated to determine reliability. These are Cronbach’s alpha, WLE (Weighted Likelihood Estimation) based reliability, and EAP/PV (Expected A Posteriori/Plausible Value) based reliability. As mentioned earlier in Chapter 2, the WLE and

EAP/PV reliabilities are based on person separation reliability. All reliability coefficients were calculated using *ConQuest*.

3.7.1.4 Model fit: Unidimensional vs multidimensional

This research postulated that the construct of AWPS is a multidimensional construct that has five separate strands. In order to check whether unidimensional or multidimensional model is a better fit for the empirical data a model comparison was conducted using *ConQuest*. When comparing model fits, two separate models, “can be evaluated based on changes in model deviance and number of parameters estimated” (Osteen, 2010, p. 77). To do this, a chi-square statistic is calculated as the difference in deviance between two models with df equal to the difference in number of parameters for the models. “A statistically significant result indicates a difference in model fit. When a difference in fit is found, the model with the smallest deviance is selected” (Osteen, 2010, p. 77). That is, the model with the smallest deviance (or the smaller deviance when only two models are compared) is considered to be a better fit. Wu and Adams (2006) also suggest using this method for comparing model fits. In addition to comparing the models, to determine the effect size of the difference between the two models, the size of the correlations between the different dimensions for the multidimensional model were considered.

3.7.2 Analysis of other quantitative data

In addition to collecting student responses to the AWPS test items developed for this research, students’ responses were collected on textbook AWPS items (which were integrated within the AWPS test). Students’ Bangla and Mathematics test scores from

school examinations were collected from the schools. These data were collected in order to establish validity evidence based on relations with other variables. Pearson correlation coefficients were calculated using SPSS to determine how the AWPS test items correlate with the textbook items and students school tests. The scaled scores from for the AWPS test were utilised for calculating this correlation.

Experts' opinion was used as a valuable resource during the test development process. In addition, an expert opinionnaire was used to collect quantitative data on the finally administered items to collect validity evidence based on test content. As the sample size for collecting expert opinion data was quite small, it was not possible to run any inferential statistical analysis on the data. Their responses were analysed with descriptive statistics, calculating percentages of response (Crocker & Algina, 2008).

3.7.3 Analysis of qualitative data from students' cognitive interviews

Various researchers and authors have discussed different methods for analysing verbal response data that has been collected for understanding question response processes. Three distinct approaches are discernible within these methods: i) quantitative, ii) qualitative, and iii) a combination of these two.

Among the researchers who have used quantitative methods, Leighton and Gokiart (2008) coded their think-aloud interview data and used frequency analysis; Leighton et al. (2011) coded and categorised their think-aloud interviews and used descriptive statistics, hierarchical linear regression, ANOVA, and item statistics for analysis. Similarly, Gadermann et al. (2011) coded and categorised their think-aloud

interviews and used frequencies, the McNemar test, Poisson or binary logistic regression analyses, and valence tests for analysis. Linninger et al. (2015) also used descriptive statistics after coding their cognitive interviews (think-aloud interviews). In addition, they employed Kruskal-Wallis tests in SPSS and Mann-Whitney tests.

Johnstone et al. (2006) used a combination of quantitative and qualitative methods for analysing their think-aloud data (captured on video). For the quantitative analysis they coded the data, calculated frequency of codes and produced descriptive statistical summaries from those. For the qualitative analysis, “all direct quotes from the students were transcribed and coded into generalized themes that represented actual events” (p. 8).

Willis (2005) outlined a qualitative method for analysing cognitive interviews which he labelled as informal. He advised on two types of analysis, one where a separate report is written for each interview to maintain case history integrity and another where the report accumulates data across different interviews providing summarised and question-oriented results. The informal analysis of cognitive interviews that he proposed has three steps: reviewing and documenting the individual interview, compiling results across interviews, and writing an organised test report. Miller et al. (2014) delineated a more elaborate qualitative method for analysing cognitive interview data. They suggested that, similar to analysing any qualitative data, the process of analysing data from cognitive interviews encompasses reduction and synthesis. They outlined five incremental steps for this reduction and synthesis. These are - “conducting interviews, producing summaries, comparing across respondents, comparing across subgroups of respondents, and reaching conclusions”

(p. 36). Although these steps are depicted individually and are presented in a linear fashion, the authors comment that “in practice they are iterative; varying levels of analysis typically occur throughout the qualitative research process” (p. 37). During the literature review for understanding analysis methods for cognitive interviews, the most detailed discussion of qualitative methods for analysing cognitive interviews was found in Willis (2015). He mentioned that the models of analysis he presented in this book were “based on a review of many cognitive testing reports and publications” (p. 58). He described five separate models of analysis for cognitive interviews - “i) text summary, ii) cognitive coding, iii) question feature coding, iv) theme coding, and v) pattern coding” (pp. 55 – 56).

After reviewing the different approaches and models of analysing cognitive interviews, it was decided that a qualitative method would be used for analysing data. There are several reasons for opting for a qualitative approach instead of quantitative or a combination of quantitative-qualitative approaches. All the reviewed studies that have used quantitative or combination methods had a large sample size for the cognitive interviews. This size ranged from 46 (Linninger et al., 2015) to 264 (Leighton et al., 2011). Given the small sample size for cognitive data collection in this research (20), it was felt that it might be difficult to generate meaningful results through quantitative analysis of the interviews in this study.

Among the qualitative approaches, the ‘reduction and synthesis’ method by Miller et al. (2014) was preferred. This method has a number of common elements with the text summary model discussed by Willis (2015).

Miller et al. (2014) extensively discussed this analysis method in their chapter. The key points they have presented for each of the steps of this analysis method are briefly discussed here:

Step 1: Conducting the interview

Analysis starts with the interview where the interviewer makes sure all relevant information has been collected and resolves any contradictions or gaps within the responses. If necessary, the interviewer intentionally redirects the respondent to make sure the targeted line of enquiry is followed.

Step 2: Producing interview summaries

Interview summaries consist of interviewers' notes about how the respondents answered the question, what experiences/perceptions they had drawn upon while answering the question, and what type of difficulties they experienced while answering the question. Any confusion or contradictions that occurred during the respondent's answering process are documented. The interview summaries give the interviewer an opportunity to systematically reflect on what occurred during the interview and make notes of his/her preliminary analysis of that.

Step 3: Developing a thematic schema

Once all the interviews are concluded, the analyst starts reviewing the respondent narratives with an intention of finding common themes across responses for the same question. These themes or interpretive patterns must be directly extracted from the interviews. Steps 1 to 3 often take place simultaneously as the themes start developing during the interviews or while summarising the interviews.

Step 4: Developing an advance schema

This step involves conducting cross-group comparison of the themes that have been identified. The analyst compares the themes to determine if any particular theme is more common for a specific respondent group. This comparison makes it possible to determine why respondents have answered the questions the way they have.

Respondent groups can be conceptualised based on the content of the questions that are being studied or based on demographic variables (e.g., gender, age, education, etc.).

Step 5: Making conclusions

The final step in this analysis process requires the analyst to put together the schemas developed earlier to determine how the questions have performed. This would include discussion on how the question was interpreted by the respondents, how responses were formulated, if they faced any difficulty in answering the question, and if there were any groups that responded to the question differently from others. These findings must be discerned from the data and preferably give an explanation of why respondents experienced difficulty and why differences occurred among different response groups.

The analysis process described by Miller et al. (2014) was used to analyse the cognitive interview data collected in the final phase of data collection in this research. Items from the same strand of AWPS were analysed together. These analyses are discussed in the 'Analysis of qualitative data' chapter.

Some post-test interviews were conducted in the pilot phase of the study as well. However, it was not possible to analyse those using the same method because of

time constraints in the field. Besides, those interviews were also a trial run of how to conduct cognitive interviews properly. An informal analysis of the students' responses in those interviews was conducted to identify item performances and any problematic items. Data from those interviews were analysed in conjunction with quantitative data from the piloting to decide on item revisions. The analysis and findings of the pilot interviews are discussed in the 'Data analysis for pilot' chapter.

3.8 Ethical considerations of this research

This research conformed to the ethical guidelines outlined by the Melbourne Graduate School of Education. The participants were informed about the process in which they were to be engaged, why their participation was required, how the data they provide would be used and to whom and how the findings would be reported. They were provided with sufficient understanding of the objectives and procedures of data collection, as well as all aspects of the research which might influence their willingness to participate in the study. Moreover, it was clarified to them during data collection that their participation was entirely voluntary and they may withdraw from the data collection any time up till the publication of the findings if they wished. They were also informed about how the data would be used confidentially and without identifying them. In addition, they were given the opportunity to get responses to their queries regarding the study after the data collection was finished.

Chapter 4: The pilot study

This chapter discusses the pilot study that was conducted as part of this research. This was a small-scale endeavour to collect data about the items (which were developed based on the AWPS framework) in order to make revisions before the larger-scale data collection is conducted. This chapter briefly recapitulates the process of item development and data collection (which have been discussed in Chapter 3 in more detail) in the beginning. Then the item analysis of the pilot AWPS test and test reliability calculations are depicted. Next, analysis of the pilot post-test interviews is described. The section after that outlines how validity evidence based on relations with other variables (i.e., students' scores in textbook items in the test and school mathematics and Bangla scores) were analysed. Finally, item revisions based on the pilot data are discussed.

As outlined in the previous chapter (Chapter 3), after the hypothesised AWPS framework with the five strands was drafted, a set of items was constructed representing the framework based on review of literature and review of existing items. Review of items before constructing new items included items from the SMART tests' item bank and items from textbooks and other supporting materials from both Australia and Bangladesh. Initially 10 -15 items were developed for each of the strands as the first draft of a test. These items represented the different ways of assessing the strands/areas of AWPS. The first draft included both short answer type and multiple-choice type items.

The draft items were then shared with the supervisors and subject matter experts. Several drafts were prepared and reviewed by the expert panel which finally

culminated in a selection of 34 items for the pilot study. This selection had six items for each of the strands of AWPS except for 'Verifying the solution and communicating the results' which had eight items. In addition to these, six items were chosen from Bangladeshi grade 4 and grade 5 textbooks for collecting validity evidence based on relations with other variables. After including the textbook items, the test for piloting had 40 items in total. Among these, 34 items were multiple choice type and six were open ended. 32 of the multiple-choice type items had five alternatives and the other (item 39 and item 40) had an open-ended response option in addition to the five alternatives. The six textbook items were also presented in multiple-choice format. All the items were designed to be scored dichotomously (i.e., correct/incorrect). So, the outcome space would be 1 for a correct answer and 0 for an incorrect answer. The open-ended items were also scored as 1 for correct and 0 for incorrect.

The pilot test was administered in one school in the Narayanganj district. In total, 62 students took the test – 33 from grade 4 and 29 from grade 5. In addition, six students from the same school were interviewed, three of them were from grade 4 and three were from grade 5. The post-test interviews were conducted in the school on the same week of the test. Data from the test were entered in SPSS by the researcher in a template that was prepared based on the test booklets earlier. Furthermore, a 10% error check was conducted to make sure the entered data matched students' responses. The SPSS data file was used for running descriptive analysis in SPSS and for doing item analysis in *ConQuest*.

4.1 Data analysis for the pilot test

Item data from the pilot test were analysed using *ConQuest*. Rasch model for dichotomous analysis was used for item analysis. However, before running the Rasch analysis, a frequency analysis was run with SPSS to check students' response patterns.

Based on the frequency analysis and students' post-test interviews, two items were excluded from the Rasch analysis. These are items 39 and 40 which were designed to assess students' use of different strategies while solving arithmetic word problems. Both had more than one correct answer, but frequency analysis showed that most of the students had chosen columnar addition/subtraction as their answer, which is one of the correct answers. In addition, students in the post-test interviews said they found these two items problematic and some had guessed an answer. These were the last two items in the test and frequency analysis showed that for these items, percentages of missing response were the highest (29% for Item 39, 27.4% for Item 40). Percentages missing for these were more than double compared to the next items with the most missing responses (12.9% for Item 38). Figure 4.1 presents the percentages of missing responses for each item. Items not mentioned here did not have any missing responses.

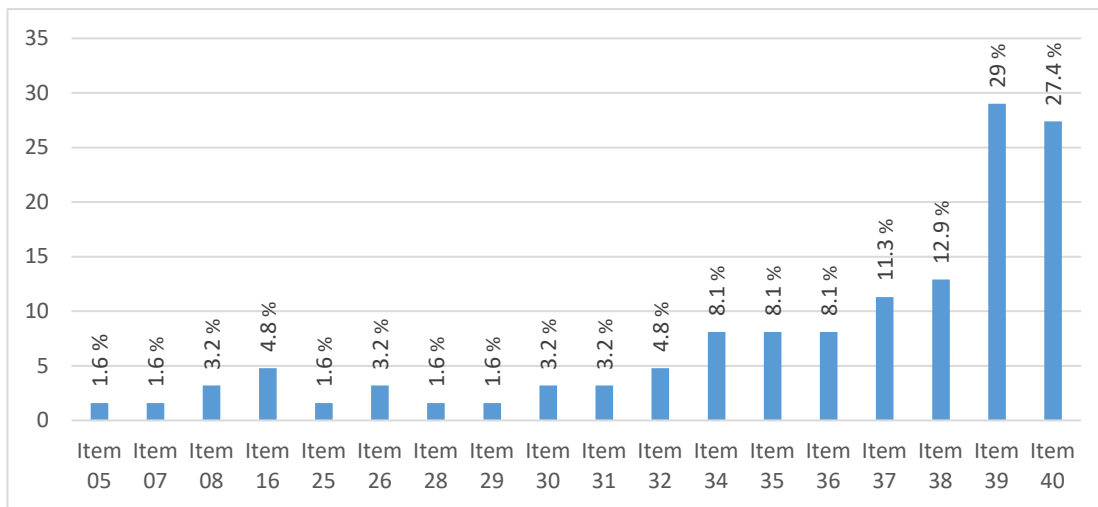


Figure 4.1. Missing response percentage per item for the pilot study

As can be seen from Figure 4.1, for most of the items, there were not a lot of missing responses. 22 items did not have any missing response at all and among the 18 items that had missing responses, more than half had less than 5% missing. If we exclude the two items with the most missing responses from the item analysis, the other five items have 8 to 13% missing responses. As the respondents were given enough time to finish the test, the possibility of them missing an item because they could not attempt it for not having enough time was quite low. Keeping this in mind, missing responses were considered as incorrect (i.e., scored as 0) while coding for item analysis in *ConQuest*.

The textbook items were excluded from the item analysis as they were part of the test to collect validity evidence. So, the Rasch analysis for the pilot test was conducted with 32 items.

4.1.1 Item analysis

This section discusses the item analysis for the pilot test.

The Wright Map

The Wright map for item analysis (presented in Figure 4.2) shows a distribution of students and items. From the Wright map it can be said that the test is portraying a reasonably good picture as an ability test. The test is fairly easy for the students. The item difficulties are well distributed, ranging from around -3 logits to 2.5 logits. There are several items at the lower end which almost all students have a probability of answering correctly and even the most able student could not correctly answer all items.

The left side of the Wright map represents students' location showing most of them are between -1 logits to 2 logits. The location map of the students resembles a bell-shaped distribution which is a common distribution found in tests measuring knowledge or ability. This could be taken as an indication of the test's suitability for assessing the students.

3									
				XX					
				XX	8 31				
				XXX					
				XX	28 35				
				XXXXXXXXXXXX	14				
2				XXXXXXX	33				
				XXXXXXXXXXXX	38				
				XXXXXXXXXXXX					
				XXXXXXXXXXXX					
				XXXXXXX					
				XXXXXXXXXXXX	22				
				XXXXXXXXXXXXXXXXXXXX	20				
				XXXXXXXXXXXXXXXXXXXXXXXXXXXX	32				
				XXXXXXXXXXXXXXXXXXXX	16				
1				XXXXXXX					
				XXXXXXXXXXXXXXXXXXXXXXXXXXXX					
				XXXXXXXXXXXXXXXXXXXXXXXXXXXX	37				
				XXXXXXXXXXXXXXXXXXXXXXXXXXXX					
				XXXXXXXXXXXXXXXXXXXXXXXXXXXX	17				
				XXXXXXXXXXXXXXXXXXXXXXXXXXXX	23 26 29				
				XXXXXXXXXXXX					
				XXXXXXXXXXXXXXXXXXXXXXXXXXXX	2 19				
				XXXXXXXXXXXX	1				
0				XXXXXXXXXXXXXXXXXXXXXXXXXXXX	7 25				
				XXXXXXXXXXXXXXXXXXXXXXXXXXXX					
				XXXXXXXXXXXXXXXXXXXXXXXXXXXX					
				XXXXXXXXXXXX	5 34				
				XXXXXXXXXXXX					
				XXXXXXX	21				
				XXXXXXXXXXXX					
				XXXXXXXXXXXX	27				
				XXXXXXXXXXXX					
-1				XXX					
				XXXXXXXXXX					
				XXXXXXX					
					11				
					15				
-2					9				
					13				
					10				
					3 4				
=====									
Each 'X' represents 0.1 cases									
=====									

Figure 4.2. The Wright map for the pilot study test (all strands together)

From the Wright map we can see that items 8 and 31 are the most difficult items and items 3 and 4 are the easiest ones.

- Item 8:** Latif has 138 plants in his house. Mitu has 112 plants more than him. Which of these number sentences would be appropriate for calculating how many plants Mitu has?
- a. $138 - 12$
 - b. $112 - 138$
 - c. $140 - 12$
 - d. $140 - 138$
 - e. None of the above (*answer*)

Item 8 is shown here. Though it was designed as an easy item assessing students' ability to map conceptual relations to numerical representations, this seemed quite hard for them to answer correctly. One reason for their difficulty could be that the correct answer for this item was 'E', that is, 'none of the above'. This is the first item in the pilot test booklet where 'none of the above' is the correct answer. This might have caused the students to choose another alternative instead of 'none of the above' because they thought they would be wrong to select 'none of the above'. Moreover, the MCQ item format that they are familiar with usually contains four alternatives instead of five, which might have made them habituated to look within the first four alternatives for an answer. Besides, some of the alternatives had the numbers mentioned in the stem which usually attracts students' attention to pick them as the correct answer.

Item 31: Water from a swimming pool has some salt in it. Tap water has the least salt. But there is much salt in water from the Atlantic or the Pacific ocean. Which water contains the most salt?

- a. Tap water
- b. Swimming pool water
- c. Water from Atlantic
- d. Water from the Pacific
- e. Not enough information to tell (answer)

Item 31 was constructed as a hard item assessing students' ability to read and comprehend a problem. The correct answer for this item is also 'not enough information to tell'. This could be difficult for the students for the abovementioned reason of students not being familiar with items not having the correct answer. But this could also be interpreted as the item working as it was intended to (i.e., a hard item) and checking students' ability to comprehend the information given to them in the stem.

Item 3: Solve this addition $24 + 31$. Show your work.

Item 4: $22 + 7 = ?$

- a. 27
- b. 24
- c. 29 (Answer)
- d. 30
- e. 15
- f. None of the above

Now if we look at the easiest items on the Wright map, items 3 and 4 were designed to be easy. Item 3 represents the easiest item for assessing students' ability to use strategies in solving AWPS. It is a simple addition without any carrying.

Similarly, item 4 was the easiest item for calculation. It was a simple addition of two numbers with the largest being a number with 2 digits. From the Wright map it seems all students were located higher than these two items and had a probability of answering these correctly.

Figure 4.3 shows the Wright map with the strands separated for a better understanding. This gives us an opportunity to compare how items from the different strands were performing. If we look at the spread of items for each of the strands, they have a good distribution of easy and difficult items. However, some strands have a better spread than others. In addition, it is noticeable how the hypothesised item

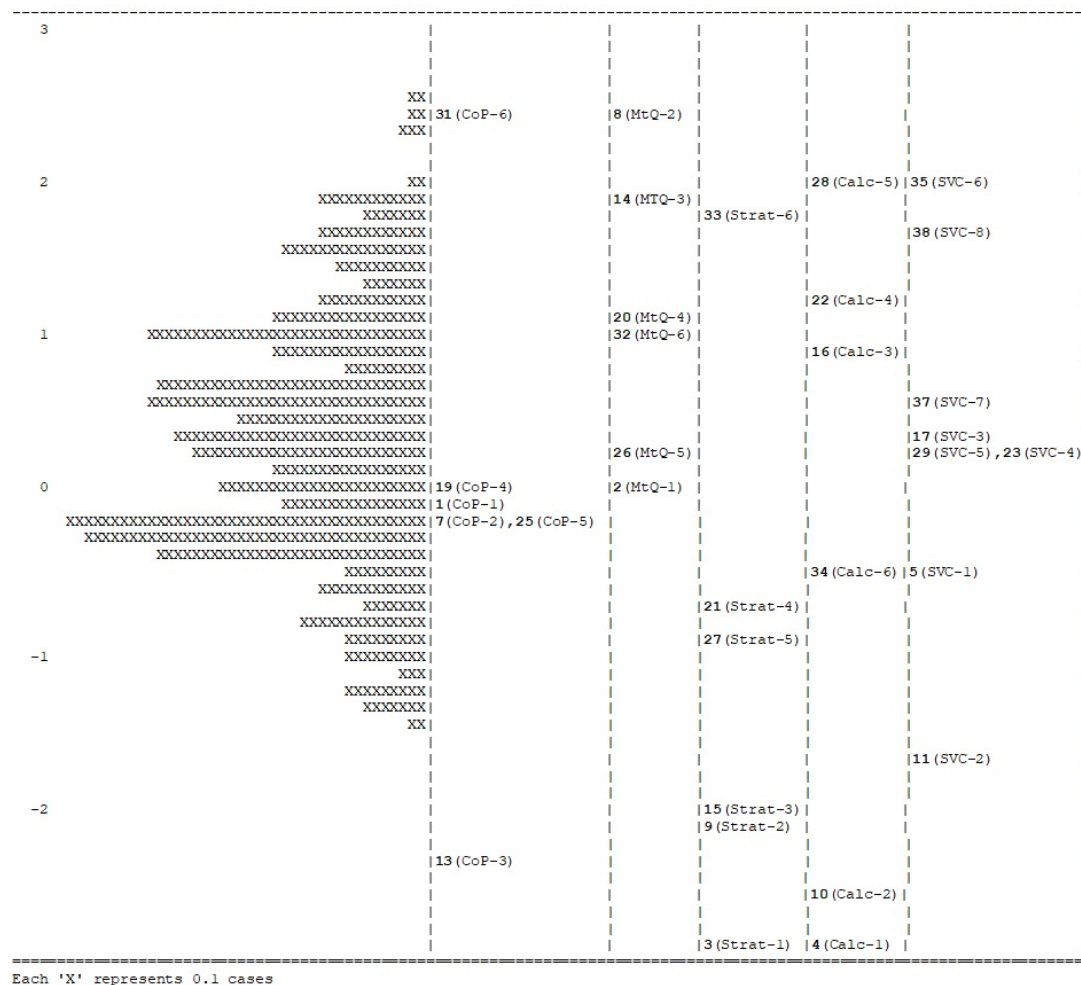


Figure 4.3. The Wright map for the pilot study test (strands separated)

difficulties were not matched in some cases. This section briefly discusses the divided Wright map for each strand.

Comprehension of the problem

Items for this strand are shown in the first column in the Wright map and are designated as CoP items. Item 13 was the easiest item for this strand. It is located at lower than -2 logits and all students have a high probability of answering this item correctly. This item (shown below) was designed to be a medium difficulty item. As the stem was quite short, the students might have easily understood the information given here, making it an easy item. It might be a good idea to change the item to increase its difficulty.

Item 13: Jamal has 5 toys, Bakar has 8 toys. Which of the following is true?

- a. Jamal has more toys than Bakar
- b. Bakar has more toys than Jamal (*answer*)
- c. Bakar has less toys than Jamal
- d. Jamal and Bakar have equal numbers of toys
- e. None of the above

The Wright map shows that the most difficult item for this strand was item 31. This was designed as the most difficult item for Comprehension of the Problem so it could be considered that it is working properly. The labels of the items in this Wright map show the hypothesised/intended difficulty of the items (i.e., 1, 2 are easy; 3, 4 are of medium difficulty and 5, 6 are hard). The item positions suggest that Item 1 and Item 7, which were hypothesised to be easy items, might be of medium difficulty. On the

other hand, Item 25 which was supposed to be a hard item is also located at a medium difficulty level.

Mapping conceptual relations to a numerical representation

Items for this strand are shown in the second column of the Wright map (Figure 4.3) and are designated as MtQ (mapping to quantitative) items. Items in this strand were difficult compared to the other strands but were well spread. The easiest item here was Item 2 which is located around 0 (zero) logits. This was designed to be an easy item but about half of the students had a probability of answering this item incorrectly. The most difficult item for this strand was Item 8 which has been discussed above. Relative positions of items within this strand suggest that Item 26, which was constructed as a hard item, turned out to be easy. Similarly, Item 32 (constructed as a hard item) was of medium difficulty. On the other hand, items 8 and 14, which were respectively hypothesised to be of low and medium difficulty were actually hard items.

Selecting an appropriate strategy for solution

Items for this strand are shown in the third column of the Wright map (Figure 4.3) and are labelled as *Strat* items. These items were easy compared to the other strands and their spread was good, ranging from -3 logits to approximately 2 logits. The easiest item here was Item 3, which all the students had a probability of answering correctly. This was constructed as the easiest item for this strand. Similarly, the most difficult item here was Item 33, which was constructed as a hard item. In addition, the hypothesised difficulty of items within this strand was well reflected in the Wright

map. However, the items were too easy compared to items from the other strands in this test. It might be helpful to revise the items to change their level of difficulty.

Performing the calculations on the stored numbers

The fourth column in the Wright map (Figure 4.3) shows the items for this strand.

These items are labelled as *Calc* (calculation) items. Items here range from -3 logits to 2 logits and are quite well spread. The easiest item here was Item 4, which was designed to be an easy item and the most difficult item was Item 28, which was designed to be a hard item. The hypothesised difficulties of items were well reflected in the Wright map except for item 34. This was designed as a hard item but turned out to be of medium difficulty. Overall, items from this strand also seem quite easy and revision of items to change their difficulty could be worth considering.

Verifying the solution and communicating the results

The final column of the Wright map in Figure 4.3 shows the items for verifying the solution and communicating the results strand. Items are labelled as *SVC* here. The item spread was good with the easiest item close to -2 logits and the hardest item at 2 logits. The easiest item here was Item 11, which was constructed as an easy item for this strand. The most difficult item here was Item 35, which was constructed as a hard item. The spread of difficulty of the items within this strand loosely follows the hypothesised difficulty levels, except for Item 29. This was constructed as a hard item but in the Wright map it is at the medium difficulty level. Similarly, items 37 and 38 were also closer to medium difficulty levels.

4.1.2 Item fit

Another crucial statistical outcome that we need to look at to understand the items' suitability is the item fit data. There are different types of fit statistics. This analysis focused on one type of fit statistics – the residual-based fit statistics. As mentioned in Chapter 3, the weighted mean-square or the infit data provides valuable insights about how individual items in a test fit the Rasch model. The fit mean-square value has an expected value of one. When an item shows misfit, there are two possibilities: if the mean-square value is less than one, it is considered as an over-fit, and when the mean-square value is greater than one, it is considered as an under-fit. The over fitting items are highly discriminating whereas the under fitting items are less discriminating (Wu et al., 2016).

If we look at Figure 4.4, the infit data for the items show that the weighted MNSQ ranges from 0.84 for item 28 to 1.15 for item 35. All the values of the weighted MNSQs are within 0.80 to 1.20. Keeping in mind the sample size was small (62 respondents) for this analysis, it could be said that all the weighted MNSQs are within a reasonable range and this does not show any extreme over-fit or under-fit. Wu et al. (2016) suggest that fit statistics should only be used to examine items so misfit items can be detected, and sources of misfit be probed. It should not be used as concrete evidence for accepting or rejecting items. They further advise that if residual-fit statistics show that items fit the Rasch model, reliability of the test items and item discrimination indices also need to be taken into account while making an overall assessment about the items' suitability.

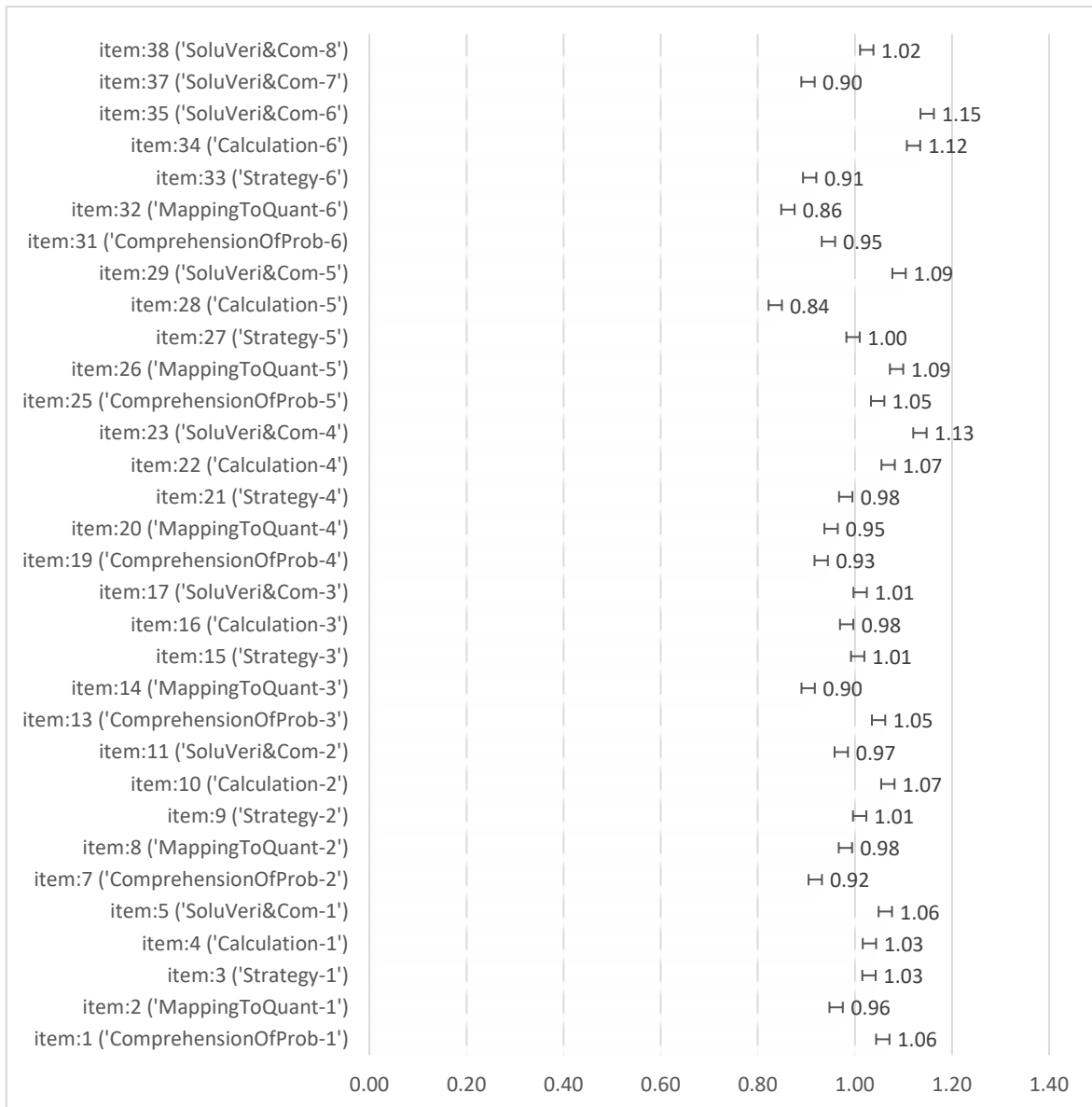


Figure 4.4. Weighted MNSQ of pilot study items

Another statistic that helps in determining item misfits is the fit t statistic. It is a transformation of the fit mean-square value accounting for the mean and variance of the fit mean-square statistic. That is, sample size is considered while doing the calculation and the resultant t statistic is a normal deviate with a mean of zero and a standard deviation of one. At the 95% confidence level, a t value outside the range of “-1.96 to 1.96” is considered as an indication of misfit. Table 4.1 presents the fit t

statistic for the items. It shows that the item fit t statistics ranges from -1.3 for item 32 to 1.5 for item 23. From these data it could be considered that the items fit the Rasch model well.

Table 4.1

Item fit t Statistics for the Pilot Study

Item	MNSQ	Confidence Interval		t
Item 1 ('ComprehensionOfProb-1')	1.06	0.81	1.19	0.6
Item 2 ('MappingToQuant-1')	0.96	0.82	1.18	-0.4
Item 3 ('Strategy-1')	1.03	0	2.3	0.3
Item 4 ('Calculation-1')	1.03	0	2.29	0.3
Item 5 ('SoluVeri&Com-1')	1.06	0.77	1.23	0.6
Item 7 ('ComprehensionOfProb-2')	0.92	0.8	1.2	-0.8
Item 8 ('MappingToQuant-2')	0.98	0.48	1.52	0
Item 9 ('Strategy-2')	1.01	0.35	1.65	0.1
Item 10 ('Calculation-2')	1.07	0.15	1.85	0.3
Item 11 ('SoluVeri&Com-2')	0.97	0.51	1.49	-0.1
Item 13 ('ComprehensionOfProb-3')	1.05	0.26	1.74	0.2
Item 14 ('MappingToQuant-3')	0.9	0.62	1.38	-0.5
Item 15 ('Strategy-3')	1.01	0.41	1.59	0.1
Item 16 ('Calculation-3')	0.98	0.8	1.2	-0.2
Item 17 ('SoluVeri&Com-3')	1.01	0.83	1.17	0.1
Item 19 ('ComprehensionOfProb-4')	0.93	0.82	1.18	-0.8
Item 20 ('MappingToQuant-4')	0.95	0.78	1.22	-0.4
Item 21 ('Strategy-4')	0.98	0.73	1.27	-0.1
Item 22 ('Calculation-4')	1.07	0.76	1.24	0.6
Item 23 ('SoluVeri&Com-4')	1.13	0.83	1.17	1.5
Item 25 ('ComprehensionOfProb-5')	1.05	0.81	1.19	0.5

Item 26 ('MappingToQuant-5')	1.09	0.83	1.17	1
Item 27 ('Strategy-5')	1	0.68	1.32	0
Item 28 ('Calculation-5')	0.84	0.6	1.4	-0.8
Item 29 ('SoluVeri&Com-5')	1.09	0.83	1.17	1
Item 31 ('ComprehensionOfProb-6')	0.95	0.48	1.52	-0.1
Item 32 ('MappingToQuant-6')	0.86	0.79	1.21	-1.3
Item 33 ('Strategy-6')	0.91	0.65	1.35	-0.5
Item 34 ('Calculation-6')	1.12	0.76	1.24	1
Item 35 ('SoluVeri&Com-6')	1.15	0.6	1.4	0.8
Item 37 ('SoluVeri&Com-7')	0.9	0.82	1.18	-1.1
Item 38 ('SoluVeri&Com-8')	1.02	0.67	1.33	0.2

4.1.3 Response fit

While answering test items, students often get distracted or for some reason do not answer the items as intended. This results in unusual response patterns and adversely affects the model fit. To identify respondents who had unusual response patterns in this test, let us take a look at the Weighted Likelihood Estimate (WLE) data for person fit. Table 4.2 shows the two lowest and the two highest person fit statistics obtained through *ConQuest*.

Table 4.2

Person Fit Statistics for the Pilot Study

PID	Students' Research ID (SRID)	Raw score	WLE	Standard Error	Fit
29	14030	11	-0.693379901	0.462054619	2.493468435
60	15028	8	-1.381834443	0.50484546	2.197185068
47	15015	24	1.793206581	0.468699225	0.405250539
56	15024	24	1.793206581	0.468699225	0.405250539

Student 29 (SRID 14030) had the largest fit value of 2.49 and this response is an underfit for the model. After sorting the student response file based on the revised item difficulties determined empirically from the Rasch analysis here, the students' response seems chaotic (Table 4.3). He/she has incorrectly answered a number of easy items but was able to answer some more difficult items correctly. The unsorted file was also checked to make sure it was not just a case of the student attempting the first few items and not answering the ones that come later. He/she seemed to have attempted all items, but the response pattern was haphazard. It is likely that he/she had randomly guessed some of the more difficult items correctly.

Table 4.3

Response Pattern for Respondent 29 (SRID 14030) when Items are Sorted by Empirically Attained Item Difficulty

Item	3	4	10	13	9	15	11	27	21
Response	0	0	0	1	1	0	0	0	0
Item	34	5	7	25	1	2	19	26	23
Response	1	1	0	1	0	0	1	1	1
Item	29	17	37	16	32	20	22	38	33
Response	0	1	0	0	0	0	1	1	0
Item	14	28	35	8	31				
Response	0	0	0	0	0				

Student 60 (SRID 15028) with the second largest fit value of 2.20 also underfits the model. In the student response file with the revised difficulty, this student answered the first five items correctly then could not correctly answer a number of easy items but correctly answered some difficult items, such as item 25 and item 8 (Table 4.4). Item 8 was discussed earlier where it was mentioned that this could have been

difficult for students because they did not expect 'none of the above' as an answer.

This student might have figured it out and answered it correctly. But still it is not clear how he/she had answered item 25 correctly. The student might have correctly guessed the answer for some of the more difficult items.

Table 4.4

Response Pattern for Respondent 60 (SRID 15028) when Items are Sorted by Empirically Attained Item Difficulty

Item	3	4	10	13	9	15	11	27	21
Response	1	1	1	1	1	0	0	0	0
Item	34	5	7	25	1	2	19	26	23
Response	1	0	0	1	0	0	0	0	0
Item	29	17	37	16	32	20	22	38	33
Response	0	0	0	0	0	0	0	0	0
Item	14	28	35	8	31				
Response	0	0	0	1	0				

Student 47 (SRID 15015) and Student 56 (SRID 15024) (responses shown in Table 4.5) both had the same response pattern, and both had the lowest fit value of 0.41 (when rounded) for the fit statistics, that is, they overfit the model. Their response patterns are consistent with the difficulties found by Rasch analysis. They show a string of 1s, that is, correct answers and then a string of 0s. Only once have they incorrectly answered an item and scored 0 within their string of 1s. That was for item 17. This might be due to a lapse of concentration. Both students have the exact same response patterns which might be an indication that one had copied the answers from the other.

Table 4.5

Response Pattern for Respondent 47 (SRID 14015) and student 56 (SRID 14024) when Items are Sorted by Empirically Attained Item Difficulty

Item	3	4	10	13	9	15	11	27	21
Response	1	1	1	1	1	1	1	1	1
Item	34	5	7	25	1	2	19	26	23
Response	1	1	1	1	1	1	1	1	1
Item	29	17	37	16	32	20	22	38	33
Response	1	0	1	1	1	1	1	0	0
Item	14	28	35	8	31				
Response	0	0	0	0	0				

Overall, the person fit statistics do not show any extreme values of person fit, apart from student 29 and Student 60 mentioned here. Response fit data for all the students are shown in Appendix 4.1.

4.1.4 Item-total correlation

Table 4.6 shows the item rest correlation data of the items in the pilot test calculated using *ConQuest*. This gives us a good indication of how the items have contributed to the overall test. Items with low or negative item-rest correlation need to be revisited and revised.

Table 4.6

Item-Rest Correlation of Items in the Pilot Study Test

Item (Sorted by Item-rest Correlation)	N	Item-Rest Correlation
Item 35 ('SoluVeri&Com-6')	62	0.00
Item 23 ('SoluVeri&Com-4')	62	0.12

Item 13 ('ComprehensionOfProb-3')	62	0.13
Item 10 ('Calculation-2')	62	0.14
Item 22 ('Calculation-4')	62	0.14
Item 3 ('Strategy-1')	62	0.16
Item 5 ('SoluVeri&Com-1')	62	0.17
Item 4 ('Calculation-1')	62	0.18
Item 8 ('MappingToQuant-2')	62	0.19
Item 34 ('Calculation-6')	62	0.19
Item 26 ('MappingToQuant-5')	62	0.20
Item 29 ('SoluVeri&Com-5')	62	0.20
Item 1 ('ComprehensionOfProb-1')	62	0.21
Item 38 ('SoluVeri&Com-8')	62	0.22
Item 25 ('ComprehensionOfProb-5')	62	0.23
Item 15 ('Strategy-3')	62	0.25
Item 27 ('Strategy-5')	62	0.26
Item 17 ('SoluVeri&Com-3')	62	0.28
Item 9 ('Strategy-2')	62	0.29
Item 31 ('ComprehensionOfProb-6')	62	0.29
Item 11 ('SoluVeri&Com-2')	62	0.30
Item 21 ('Strategy-4')	62	0.32
Item 16 ('Calculation-3')	62	0.33
Item 2 ('MappingToQuant-1')	62	0.36
Item 20 ('MappingToQuant-4')	62	0.37
Item 19 ('ComprehensionOfProb-4')	62	0.40
Item 33 ('Strategy-6')	62	0.41
Item 37 ('SoluVeri&Com-7')	62	0.42
Item 14 ('MappingToQuant-3')	62	0.43
Item 7 ('ComprehensionOfProb-2')	62	0.47
Item 28 ('Calculation-5')	62	0.50
Item 32 ('MappingToQuant-6')	62	0.51

From Table 4.6 we can see that Item 35 has a zero correlation. This item is not contributing to the overall test, that is, student response on this item is not correlated well with the rest of the items in the test. Students might have had trouble understanding what the item wanted them to do and they are likely to answer it randomly. It would be helpful to revise or delete the item.

Item 35: The headteacher of a school is planning a victory day celebration in her school and wanted each student to have a small Bangladesh flag. These flags come in packets of 200s. The school has 656 students in total but 126 of them will not be present. The headteacher calculated there will be 530 students in the celebration and the school has 4 packets of flag. Which of the following alternatives show the correct calculation and result?

- a. $656 + 126 = 530$; 4 packets of flag
- b. $656 - 126 = 530$; 4 packets of flag
- c. $656 - 126 = 530$; 3 packets of flag
- d. $656 - 126 = 530$; 2 packets of flag
- e. $656 + 126 = 530$; 3 packets of flag

Other items with low item rest correlation include item 23 (0.12), item 13 (0.13), item 10 (0.14), and item 22 (0.14). It would be good to take a closer look at these items and make amendments to make them better.

Item 23: Suppose you were asked to subtract 129 from 2613 and you found the answer was 2484. How would you write the solution down?

a. 2631

$$\begin{array}{r} 2631 \\ -129 \\ \hline 2484 \end{array}$$

b. -129

$$\begin{array}{r} 2613 \\ -129 \\ \hline 2484 \end{array}$$

c. 2613

$$\begin{array}{r} 2613 \\ -129 \\ \hline 2484 \end{array} \text{ (Answer)}$$

d. 2613

$$\begin{array}{r} 2613 \\ -129 \\ \hline 2484 \end{array}$$

e. None of the above

Item 13: Jamal has 5 toys, Bakar has 8 toys. Which of the following is true? (Difficulty: medium)

- a. Jamal has more toys than Bakar
- b. Bakar has more toys than Jamal
- c. Bakar has less toys than Jamal
- d. Jamlar and Bakar have equal numbers of toys
- e. None of the above

Item 10: $23 - 8 = ?$

- a. 15 (Answer)
- b. 31
- c. 17
- d. 16
- e. None of the above

Item 22: Which of the following is the closest to the result for solving $87 - 53$?

- a. 37 (Answer)
- b. 45
- c. 27
- d. 30
- e. None of the above

4.1.5 Analysis of test reliability

The WLE person separation reliability calculated in *ConQuest* for the pilot test was 0.77, which indicates that the test was quite sensitive in distinguishing between high and low performers.

The EAP/PV reliability calculated in *ConQuest* for the test was 0.79. Considering the test was administered with a small sample of 62 students and 32 items were considered for the analysis, the reliability statistics are quite good. In addition to these, Cronbach's Alpha, which is calculated based on raw scores, was calculated in *ConQuest* and returned a coefficient alpha value of 0.77.

4.1.6 Person standard errors

Person standard errors calculated as the Standard Error of Measurement (SEM) is an important analysis for determining the usefulness of the estimate of respondents' location (Wilson, 2005). It can provide us with valuable insights about the test's reliability. For the pilot AWPS test, students' ability estimates (WLE) and the standard errors were calculated in *ConQuest*. Figure 4.5 demonstrates how the standard errors were distributed for the different ability estimates. It shows that the ability estimates ranged between approximately -1.5 logits to 2.5 logits and the standard errors ranged from approximately 0.45 to 0.55. This suggests that the errors in measurement were low, that is, the test was reliable.

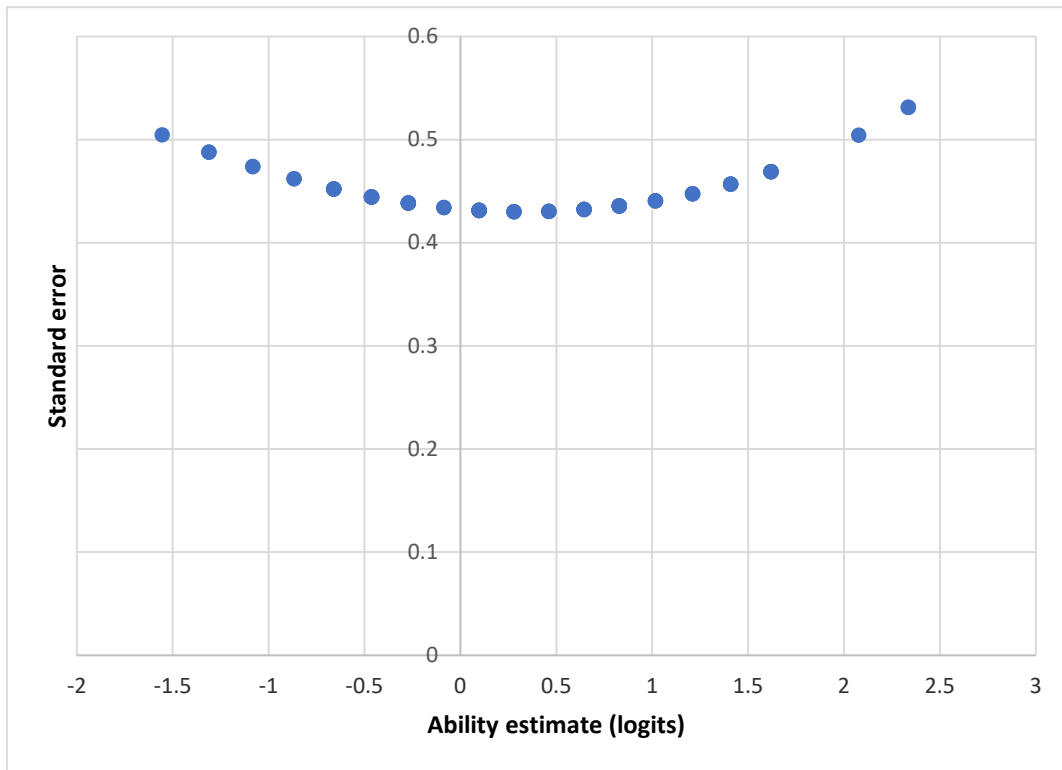


Figure 4.5. Standard error of measurement for the pilot AWPS test

4.1.7 Students' post-test cognitive interviews

Post-test cognitive interviews (i.e., think-alouds) were conducted with six students selected through convenience sampling in the pilot study. These were crucial in understanding how best to conduct the interviews and helped in developing the standard operating procedures (discussed in Chapter 3) for conducting post-test cognitive interviews in the final phase. As mentioned in the data analysis section of the previous chapter, the pilot interviews were informally analysed to determine item performances (i.e., whether the items were tapping into the areas they are meant to) and to determine any problematic items.

Overall, analysis of the interviews suggested that the students were using the skills they are meant to use for the different strands of the framework, that is, they seemed to depend on their reading ability for answering the ‘comprehension of the problem’ items; they were identifying the numbers from the word problems and how those numbers are to be related for the ‘Mapping conceptual relations to numerical representation’ items and so on. However, there seemed to be some overlap in the processes they were following for the ‘Selecting an appropriate strategy for solution’ and ‘Performing the calculations on the numbers’ items.

Item 37: Pia was asked to solve this problem:

‘Rashed started for his school from home at 10.40 am and reached the class at 11.30 am. How long did it take for him to reach his class?’

She solved the problem and found the answer was 50. How shall she write down the result?

- a. 50 am
- b. 50 hours
- c. 50 minutes
- d. 0.50 hours
- e. None of the above

Item 38: Leon was asked to solve the following problem:

‘Jenny bought 5 candies for 12.5 taka. She gave the shopkeeper a 20-taka note. How much shall she get in change?’

He solved the problem and found the answer was 7.5. How should he write down the answer?

- a. 7 taka
- b. 7 taka and 50 poisa
- c. 7.5 poisa
- d. 7.5
- e. None of the above

During the interviews, while answering items 37 and 38, the students seemed to go directly to the mathematical problems within the quotation marks and tried to find

the result. These items were presented as scenarios with answers and were meant to be verifying students' ability to determine the correct presentation of the answer. But the students did not read the setup (e.g., for item 37, "Pia was asked to solve this problem, she solved the problem and found the answer was 50. How shall she write down the result?") which could adversely affect the way these items are meant to work. It was felt that this happened partly due to the quoted item being in a separate paragraph and the scenario and answer in separate paragraphs. It might be helpful to revise these items to make the stem a single paragraph.

The students were very confused about items 39 and 40. They could not understand the items well and when the items were explained to them, they selected the typical columnar addition/subtraction strategies. The students did not seem to be familiar with any strategies, other than columnar addition/subtraction and these items were taking a long time for them to answer. So, it might be better to delete these two items.

4.2 Analysis of validity evidence based on relations with other variables

The pilot test had six traditional AWPS items from textbooks which were included to collect validity evidence based on relations with other variables. In addition, students' Mathematics and Bangla school-test scores were collected as another source of validity evidence of the same type.

To compare students' performance in the AWPS test (total 32 AWPS items) with the other assessment types (i.e., textbook items and school tests, which are typically used and generally accepted by stakeholders), students' performance in logit values obtained from the Rasch analysis were converted using a linear transformation.

This generated a scaled score (where Scaled Score = Ability estimate in logit $\times$ 100 + 500) which was used to calculate the correlation between students' performance in the AWPS test and their total score in textbook items. It was also used for determining correlation between the AWPS test and students' Mathematics and Bangla test performance in school-tests. These correlations were calculated in SPSS.

Pearson correlation between students' scaled score in the AWPS test and their total score in the textbook items was 0.52 (N = 62). This was statistically significant at the 0.01 level (2-tailed).

Pearson correlation between the scaled score in AWPS and Mathematics score in school test was 0.39 (N = 60) and between scaled score and Bangla score was 0.48 (N = 60). Both these correlations were also statistically significant at the 0.01 level (2-tailed).

Cohen (1988) suggested that correlation coefficients in the order of 0.10 are small, in the order of 0.30 are medium and in the order of 0.50 are large.

Furthermore, based on his review of two large meta-analysis reviews which were conducted previously, Hemphill (2003) classified the correlation coefficients found in those studies into three equal groups. He found that the correlation coefficients for the upper third of the studies had a coefficient value of >0.30 . That is, two-thirds of the studies (those in the lower third and the middle third of the distribution) had a correlation coefficient up to 0.30. Even though the sample number for the analyses conducted in the pilot study was small, the correlation values found here can be considered as good based on the criteria discussed by Cohen (1988) and Hemphill (2003). The correlations show there is evidence of validity for the AWPS test.

4.3 Item revision based on pilot data

The pilot study suggested that most of the strategy items had a low average delta, that is, they were easy items (Table 4.7). This is visible in the Wright map too where they are distributed from -3 logits to about 2 logits.

Table 4.7

Item Difficulty of Strategy Selection Items in the Pilot Study Test

Item (Sorted by avg delta)	N	Avg Delta
Item 3 ('Strategy-1')	62	-3.21
Item 9 ('Strategy-2')	62	-1.993841865
Item 15 ('Strategy-3')	62	-1.810927803
Item 27 ('Strategy-5')	62	-0.80926162
Item 21 ('Strategy-4')	62	-0.531184203
Item 33 ('Strategy-6')	62	1.998983017

Based on the piloting data these items were changed for the final version of the test.

The item specifications for the piloting items and how they were revised for the final test are described in Table 4.8.

Table 4.8

Revising Item Specifications for the Strategy Selection Items

Item number	Item specification in the pilot study	Revised item specification for final data collection	Difficulty level
3	Addition of two two-digit numbers without carrying	Addition of two four-digit numbers using carrying once	Low
9	Subtracting two two-digit numbers without borrowing	Subtracting two four-digit numbers without borrowing	Low
15	Addition of two two-digit numbers using carrying once	Addition of two five-digit numbers using carrying twice	Medium

21	Subtracting two two-digit numbers using borrowing once	Subtracting two four-digit numbers using borrowing once	Medium
27	Addition of three numbers (four digit + three digit+ two digit) with carrying once	Addition of three five-digit numbers using carrying three times	Hard
33	Subtracting three numbers (four digits – three digits – two digits), without borrowing (adding two numbers first to subtract)	Subtracting three four-digit numbers where two smaller numbers need to be added first	Hard

While revising the items, initially the idea was to include multiple numbers of digits for a single item, that is, addition of four-digit and three-digit numbers, subtraction of four-digit numbers from three-digit numbers and so on. But to make sure that students' understanding of place value does not interfere with assessing the strategies (since these are open ended items) the idea of mixing different number of digits was abandoned. Instead, all of them were made with same number of digits but the number in the thousandth or higher places were usually made a small number like 1 or 2.

Table 4.9

Item Difficulty of the Performing Calculation Items in the Pilot Study Test

Item (Sorted by avg delta)	N	Avg Delta
Item 4 ('Calculation-1')	62	-3.21
Item 10 ('Calculation-2')	62	-2.456333271
Item 34 ('Calculation-6')	62	-0.359627607
Item 16 ('Calculation-3')	62	1.096553016
Item 22 ('Calculation-4')	62	1.425783762
Item 28 ('Calculation-5')	62	2.227922924

In addition to the strategy selection items, three items from the ‘Performing the calculations on the numbers’ were deemed too easy with low average delta (as shown in Table 4.9) and with their location in the Wright map. These items (Item 4, 10, and 34) were also revised to make them more difficult. Table 4.10 describes the specifications for these items in the pilot study and the revised specification for the final test.

Table 4.10

Revising Item Specifications for the Performing Calculation Items

Item number	Item specification in the pilot study	Revised item specification for final data collection	Difficulty level
4	Addition of two numbers (maximum two digits)	Addition of two numbers (maximum three digits)	Low
10	Subtracting two numbers (maximum two-digits)	Subtracting two numbers (maximum three-digits)	Low
34	Subtracting two numbers, maximum four digits	Subtracting two numbers, both four digits	Hard

For some of the other items (e.g., 6, 12, 20, 22, 32, 34) there were certain distractors that the students had not chosen at all or that only few students had chosen. So, some of these distractors were revised using more plausible ones.

Item 35 showed a zero item-rest correlation and it was determined the item had some problems with the phrasing in Bangla making it confusing for students. So the translation was revised.

From the post-test cognitive interviews it was found that, for items 37 and 38, the quoted questions in a separate paragraph created some issues. To alleviate this, the whole stem was formatted to be a single paragraph in the revised test.

Items 39 and 40 (these items presented different strategies as options for solving an addition/subtraction problem) seem to be the most confusing ones for the students. Many of them skipped these items or gave multiple answers. They also conveyed their discomfort with these items during the post-test interviews. Consequently, these items were omitted from the final test.

Chapter summary

This chapter presented the data analysis of the pilot study. Analysis of the pilot test (e.g., the Wright map, item fit, response fit, test reliability, etc.) was discussed. In addition to the psychometric analysis of the test, the informal analysis of the students' post-test cognitive interviews was described. Furthermore, analysis of validity evidence based on relations with other variables (e.g., the textbook items and students' mathematics and Bangla test scores from school) was outlined. The chapter concluded with a discussion on how the test items were revised based on the pilot data analysis.

Chapter 5: Analysis of quantitative data

This chapter discusses the analyses of the quantitative data that were collected in the final phase of this research. These data include the Students' AWPS test, the textbook items that were integrated within the test, students' Bangla and Mathematics test scores from their schools and data from the Expert Opinionnaire. The data analyses are discussed based on the different types of validity evidence they represent. First, the validity evidence based on the internal structure of the test is discussed: a brief recap of how the final test was constructed and administered is presented and the item analyses for the AWPS test (e.g., the Wright map, item fit, response fit, item-total correlation, etc.) are described. Next, the model fit for the AWPS test is compared between a Unidimensional and Multidimensional model. The section after that presents the reliability analyses for the AWPS test. The subsequent section presents the analysis of validity evidence based on relations with other variables. This covers comparison of students' performance in the AWPS test with their textbook item scores and their school (Bangla and Mathematics) test scores. The final section of the chapter presents the analysis of validity evidence based on test content collected in the form of an Expert Opinionnaire.

5.1 Analysis of validity evidence based on internal structure: Students' AWPS test

The initial test was piloted and revised based on the pilot data. Items in the final version of the test had the following distribution for the different cognitive

elements/strands of arithmetic word problem solving (AWPS) and textbook items (Table 5.1).

Table 5.1
Distribution of Items in the Final Test Booklet

Type of items	Number of items	Item serial number in the test booklet
Comprehension of the problem	6	1, 7, 13, 19, 25, 31
Mapping conceptual relations to a numeral representation	6	2, 8, 14, 20, 26, 32
Selecting an appropriate strategy	6	3, 9, 15, 21, 27, 33
Performing calculations on the numbers	6	4, 10, 16, 22, 28, 34)
Verifying the solution and communicating the results	8	5, 11, 17, 23, 29, 35, 37, 38
Textbook Items	6	6, 12, 18, 24, 30, 36

Except for ‘Verifying the solution and communicating the results’, which had eight items, all the other strands had six items each. In addition, there were six textbook items (which were typical AWPS items) included to collect validity evidence. In total, the test had 38 items, 32 of them representing the AWPS framework and six taken from primary textbooks.

Students’ test data were collected from nine schools in the Narayanganj district of Bangladesh. Among these, seven were Government primary schools (which is the prevalent type of primary school in Bangladesh), one was a privately managed school and the remaining one was a primary school attached to a high school. In terms of the schools’ self-reported geographical locations, two were from rural areas, four from semi-urban areas, and three from urban areas. However, it should be kept in

mind that the schools were from a single district, so the varying geographical locations might not represent an actual nationwide scenario of the specified locations. Data were collected from both grades 4 and 5 in all the participating schools. In total, 826 respondents took the test in the nine participating schools. After collecting the test booklets from the schools, the data were coded, entered in a database, and checked for errors. In this process of data cleaning, 18 respondents were removed from the final data set due to various types of incomplete data (e.g., most of the item responses missing, student identification details missing, etc.). As a result, the final data set used for analysis included 808 respondents. The following sections discuss the analyses of data from the test.

5.1.1 Frequency analysis of the AWPS test data

Similar to the pilot study, a single dimension Rasch model for dichotomous analysis was used for item analysis. Before running the Rasch analysis, a frequency analysis was run with SPSS to check students' response patterns. This was particularly helpful in deciding whether to consider missing responses as 'missing' or as 'incorrect' for the

Rasch analysis in *ConQuest*. Figure 5.1 shows the percentage of missing response for each item in the AWPS test (excluding the textbook items). The percentage missing ranges from 0.10% for Item 01 to 5.80% for Item 33. Only two items had more than 5% missing responses. Those were Item 35 (5.70%) and Item 33 (5.80%).

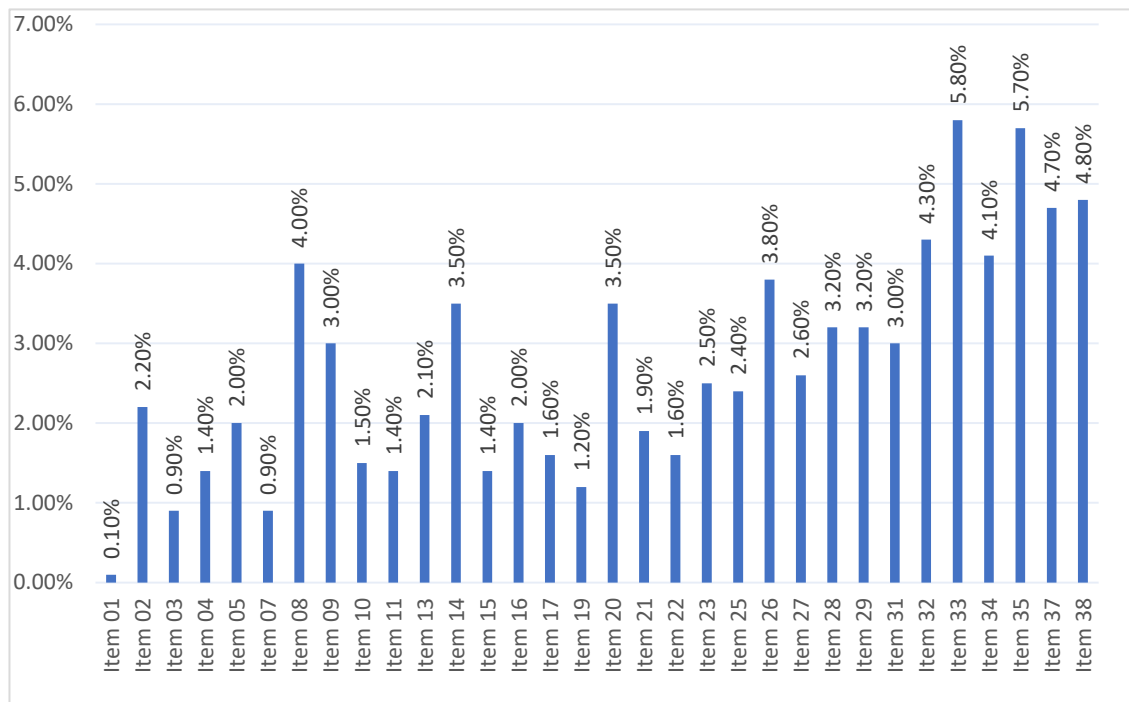


Figure 5.1. Percentages of missing response for each item (textbook items excluded)

Taking all available information into consideration, it was decided that missing responses were to be considered as incorrect in the item analysis conducted in *ConQuest*. The three main reasons for doing this are as follows:

- a) As per the ground rules decided during data coding (discussed in Chapter 3), many missing response entries were, in reality, responses where students had chosen more than one answer or crossed off all their answers. So, it would be logical to consider them as incorrect.

b) The data on missing responses in Figure 5.1 show that the items with the most missing responses were Item 33 (5.80%) and 35 (5.70%). Items that came in the booklet after these two, that is, Item 34, Items 37 and 38 had a much lower missing response percentage (4.10% for 34, 4.70% for 37 and 4.80% for 38). So, for a number of students whose responses were coded as 'missing' for these two items, these were not the last items in the booklet that they did not reach since they answered items coming after these.

c) Furthermore, based on the piloting of items, the time given to students for completing the items was considered adequate. Therefore, the possibility of them missing an item because they could not try it was quite low.

The textbook items that were part of the test booklet were excluded from the item analysis, so the Rasch analysis was conducted with 32 items.

5.1.2 Item analysis

This section discusses the item analysis for the test.

The Wright map

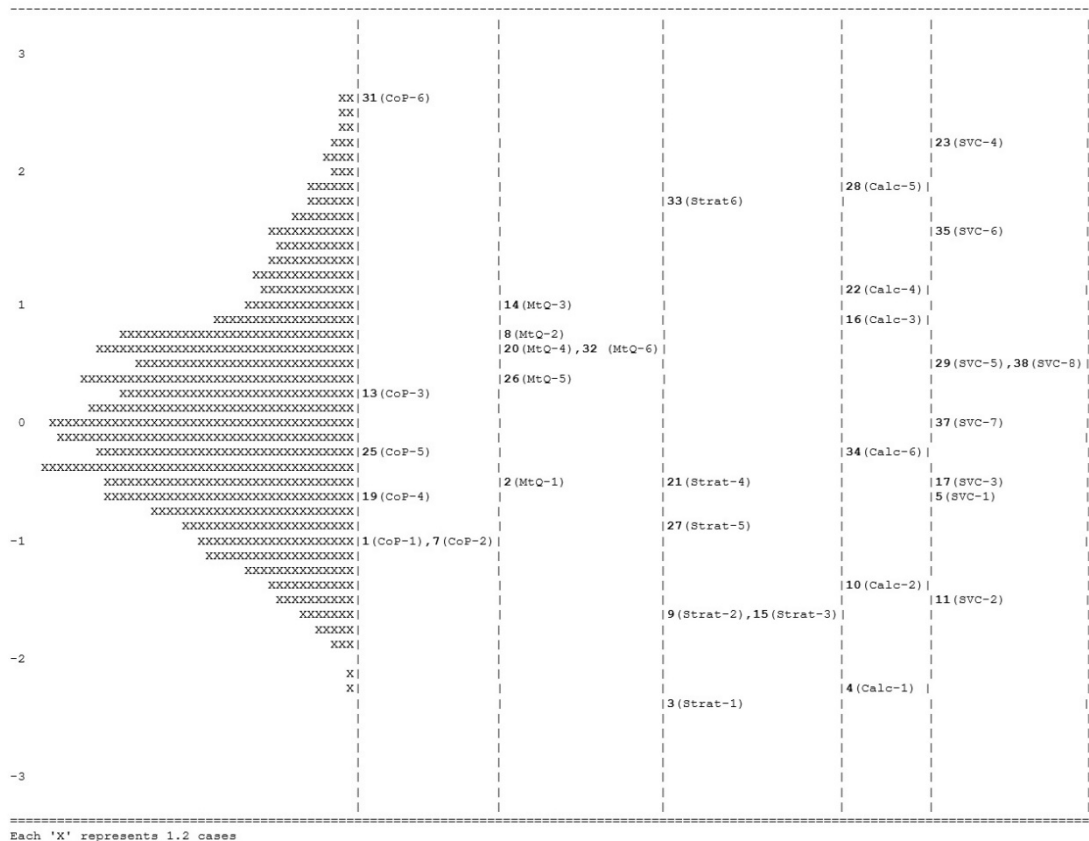


Figure 5.2. The Wright map for final data collection test (strands separated)

The Wright map (shown in Figure 5.2) for item analysis shows a distribution of students and items. From the Wright map it can be said that the items match the students' abilities quite well. The item difficulties are well distributed, ranging from approximately -2.5 logits to 2.5 logits. There are a couple of items at the lower end (items 3 and 4) which almost all students have a probability of answering correctly and

only a few very able students (2.4 cases according to the wright map) have a probability of correctly answering all items.

The left side of the Wright map represents students' location, showing most of them are between -2 logits and 2 logits. The location map of the students resembles a bell-shaped distribution which is a common distribution found in tests measuring knowledge or ability.

The Wright map shown here has the strands separated which gives us an opportunity to note and compare how items from the different strands have performed. Looking at the distribution of items in each strand, we could say they have a good spread of easy and difficult items matching the range from the lowest student location to the highest. Items in the third, fourth and fifth strands have a good dispersion in the Wright map. However, items in the first strand, Comprehension of the Problem, have an inclination toward the positive logits. The two easy items from this strand are more difficult than some of the easy and medium difficulty items in the last three strands. In addition, the most difficult item (Item 31) in the test is from this area located at approximately 2.5 logits.

Items in the second strand, Mapping Conceptual Relations to a Numerical Representation, show the least dispersion. All the items of this strand are clustered within approximately -0.50 to +1 logit. The following sections discuss each of the strands in the Wright map.

Strand 1: Comprehension of the problem

The first column of the Wright map shows the location of the items for this strand.

These are designated as CoP items. The easiest items in this strand are 1 and 7 located

at -1 logits. These were designed as the easiest items for this area and seem to be working well. However, a number of respondents who are located below -1 logits level have a high probability of not answering these items correctly. Item 13 of this strand was initially designed to be a medium difficulty item, but it was revised after the piloting as it originally turned out to be the easiest. The revised item seems to be quite difficult for the students.

Item 13: Jamal has more toys than Bakar who has more toys than Rimi but less than Rafi. Which of the following is true?

- a. Bakar has more toys than Rafi
- b. Rafi has more toys than Rimi (answer)
- c. Rimi has more toys than Rafi
- d. Rafi has more toys than Bakar (answer)
- e. None of the above

As this item requires the respondents to do some abstract reasoning, it is logical to believe it worked well as a hard item.

The most difficult item in this area is Item 31, located at approximately 2.5 logits. This was also the hardest item in the whole test.

Item 31: Water from a swimming pool has some salt in it. Tap water has the least amount of salt. But there is much salt in water from the Atlantic or the Pacific Ocean. Which water contains the most salt?

- a. Tap water
- b. Swimming pool water
- c. Water from the Atlantic
- d. Water from the Pacific
- e. Not enough information to tell

This item was designed to be a hard item for the students. It has a long stem that requires students to understand several points of information and then comprehend the fact that it is not possible to determine the single answer from the information given here. The item seemed to have worked well as a hard item. Frequency

distribution of responses for this item showed that 78.90% students chose either C or D as their answer (37.4% chose C and 41.5% chose D). C or D could be considered as reasonably good answers for this item but was not considered as the correct answer because the idea was to determine students' ability to understand the information given in the text and it is likely that they partially understood the stem while answering C or D. However, in future iteration of this test it might be worth revising this item.

Strand 2: Mapping conceptual relations to a numerical representation

Items from this strand are shown in the second column of the Wright map. These are designated as MtQ (mapping to quantitative) items. The easiest item for this strand was item 2 located at approximately -0.50 logits. It was designed as an easy item and performed well. It was difficult for several respondents who are located below this item's location and they have a high probability of answering this item incorrectly.

Item 14: Two hundred and twenty-three boys, one hundred and thirty-five girls, and twelve teachers participated in a school trip. Which of these number sentences is correct for finding out the total number of people going on the trip?

- a. $223 + 135$
- b. $135 + 12 - 223$
- c. $223 - 135 - 12$
- d. $12 + 135 + 223$ (Answer)
- e. None of the above

A number of items in this strand showed a dissimilar difficulty than what was hypothesised for them. Items 8 and 14 which were hypothesised to be easy and medium difficulty items respectively, turned out to be the hardest items. Items 26 and 32, hypothesised as the hardest items, showed medium difficulty.

Item 14 is the hardest item in this strand, located at approximate 1 logits. This was probably difficult for students because they are habituated to put down the numbers from the stem in the sequence they are mentioned. As a result, many could not map the relation among the numbers properly.

Item 8: Latif has 138 plants in his house. Mitu has 112 plants more than him. Which of these number sentences would be appropriate for calculating how many plants Mitu has?

- a. $138 - 12$
- b. $112 - 138$
- c. $140 - 12$
- d. $140 - 138$
- e. None of the above (Answer)

Item 8 was designed as an easy item. But it turned out to be the second hardest item within this strand. This might be because many students chose B as their answer, which is wrong. They picked the two numbers mentioned in the stem and without heeding the relation between these two numbers, chose an alternative that had them. Another reason for finding this item difficult could be their unfamiliarity with 'None of the above' as an answer. This has been discussed in the piloting chapter extensively.

Strand 3: Selecting an appropriate strategy for solution

The third column of the Wright map shows the items for this strand. These are labelled as *Strat* items. The items are located between approximately -2.5 logits to 2 logits and have a good dispersion. The easiest item in this area was Item 3 which was designed as an easy item. Similarly, Item 33, designed as a hard item, was the most difficult item here. Items 9 and 15, which were designed as easy and medium difficulty items, are in the middle of the Wright map in the same level. Item 9 was about subtracting two four-digit numbers without borrowing and Item 15 was about adding

two five-digit numbers with carrying twice. Item 27, designed to be a difficult item, turned out to be in the medium difficulty area. On the other hand, item 21 which was hypothesised as a medium difficulty item was found to be hard.

Item 27: Solve this addition $40695 + 17582 + 10346$. Show your work.

Item 21: Solve this subtraction $6867 - 2593$. Show your work.

Item 27 was about adding three 5-digit numbers where students had to use carrying strategies three times, whereas Item 21 was about subtracting two four-digit numbers while using borrowing strategies once. Looking at these four items (9, 15, 21, and 27) it could be suggested that students found the addition strategies (carrying over) to be easier to implement than subtraction strategies (borrowing). That is why the addition strategy items were easier for them than what was hypothesised.

Strand 4: Performing the calculations on the numbers

Items for this strand are shown in the fourth column of the Wright map. They are labelled as *Calc* (calculation) items. Items in this area show a good dispersion and are located between approximately -2.5 logits to 2 logits but they are slightly more difficult than the strategy items. The easiest item here is Item 4 and the hardest is Item 28 which are same as their hypothesised difficulty. Other item locations also reflect the difficulties they were designed for, except for Item 34.

Item 34: Which of the following is the correct answer for solving $7322 - 3980$?

- a. 3342 (Answer)
- b. 4342
- c. 3242
- d. 4242
- e. None of the above

Item 34 was hypothesised to be a hard item but turned out to be of medium difficulty.

One reason for this could be the fact that all the items that are more difficult in this strand according to the Wright map, had an additional level of calculation the respondents had to do. Those items did not provide the direct answer of the calculation as an alternative for the respondents to choose.

Item 22: Which of the following is the closest to the result for solving $87 - 53$?

- a. 37 (Answer)
- b. 44
- c. 27
- d. 30
- f. None of the above

Take item 22 as an example. After doing the subtraction, respondents would have got 34 as the result. Then they would have had to calculate which of the alternatives were the closest to 34 to find the correct answer. This probably made these three items more difficult and Item 34 easier.

Strand 5: Verifying the solution and communicating the results

The final column of the Wright map shows the items for this strand. Items are labelled as SVC (Solution verification and communication) here. The items are located between approximately -2 logits and 2 logits and have a good dispersion along the Wright map. The easiest item here is Item 11 which was designed as one of the easy items. Except for Items 23 and 37, other items also reflect the hypothesised difficulty with Item 17 (medium difficulty item) located around the middle of the Wright map and Items 29, 35 and 38 (hard items) located high in the Wright map.

Item 23: Suppose you were asked to subtract 129 from 2613 and you found the answer was 2484. How would you write the solution down?

a.
$$\begin{array}{r} 2631 \\ -129 \\ \hline 2484 \end{array}$$

b.
$$\begin{array}{r} -129 \\ 2613 \\ \hline 2884 \end{array}$$

c.
$$\begin{array}{r} 2613 \\ -192 \\ \hline 2484 \end{array}$$

d.
$$\begin{array}{r} 2613 \\ 129- \\ \hline 2484 \end{array}$$

e. None of the above (answer)

Item 23 was designed as a medium difficulty item, but the Wright map shows that it is the hardest item in this strand. Respondents were supposed to look for the correct presentation of the solution and result among the alternatives and choose the correct answer, which for this item was 'None of the above'. However, the first distractor (A) here was almost correct with only the number 2631 instead of 2613. Most respondents failed to verify the solution properly and chose A as their answer. That is

what made this item so difficult for them. Another reason might be that they do not believe 'None of the above' is a possible answer.

Item 37 was designed to assess students' ability of understanding real life situations. It was hypothesised to be a hard item but in reality, students were quite familiar with calculating times and using proper time units so this turned out to be a medium difficulty item for them.

Item 37: Pia was asked to solve this problem:

"Rashed started for his school from home at 10.40 am and reached the class at 11.30 am. How long did it take for him to reach his class?"

She solved the problem and found the answer was 50. How shall she write down the result?

- a. 50 am
- b. 50 hours
- c. 50 minutes (answer)
- d. 0.50
- e. None of the above

5.1.3 Item fit

Item fit data provides important information that helps us to understand the items' suitability. There are different types of fit statistics. Among them, we are considering the residual-based fit statistics. To recap what has been discussed previously, the weighted mean-square or the infit data provides valuable information regarding how individual items in a test fit the Rasch model. The fit mean-square value has an expected value of 1. There are two possibilities when an item shows misfit: if the mean-square value is less than one, it is considered as an over-fit, and when mean-square value is greater than one, it is considered as an under-fit. The over fitting items

are highly discriminating whereas the under fitting items are less discriminating (M. Wu et al., 2016).

Figure 5.3 shows the weighted MNSQ (infit) statistics for the items. The weighted MNSQs range from 0.88 for item 28 to 1.25 for item 35. All the other MNSQs are within 0.80 to 1.25. Considering Figure 5.3, it could be said that all the items are within a reasonable range and do not show any extreme over-fit or under-fit. Only Item 35 has a weighted MNSQ of 1.25, which could be considered as an underfit. So, it might be worth looking at this item. Wu et al. (2016) suggest that fit statistics should only be used to examine items, so misfit items can be detected, and sources of misfit can be probed. It should not be used alone as evidence for accepting or rejecting items but should be combined with evidence regarding item content.

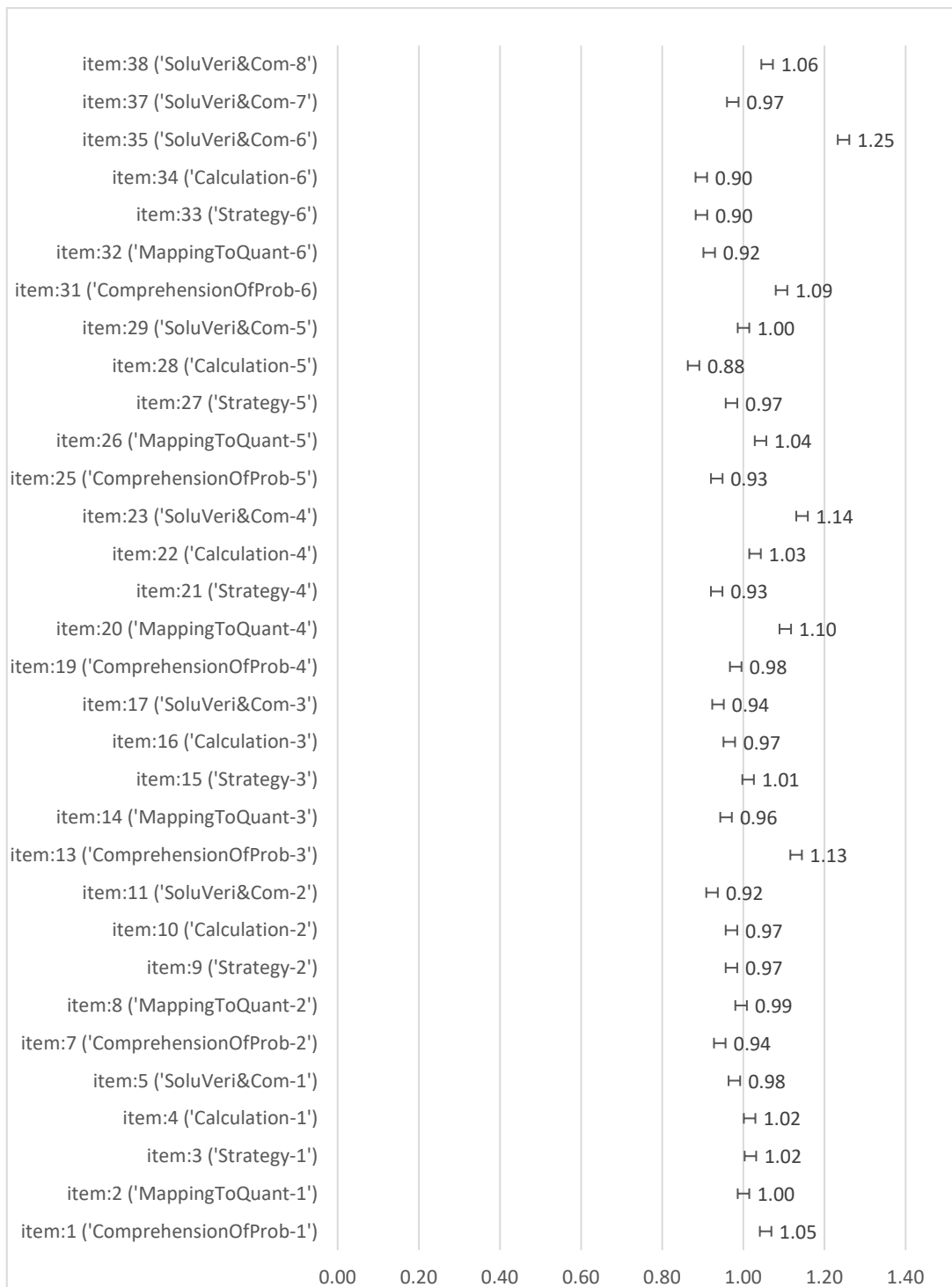


Figure 5.3. Weighted MNSQ (infit) of items

Another statistic that helps in determining item misfits is the fit t statistic. It is a transformation of the fit mean-square value accounting for the mean and variance of the fit mean-square statistic. That is, sample size is considered while doing the

calculation and the resultant t statistic is a normal deviate with a mean of zero and a standard deviation of one. At the 95% confidence level, a t value outside the range of “-1.96 to 1.96” is considered as an indication of misfit.

Table 5.2 presents the fit t statistic for the items. In total, ten items have a t value outside the range of -1.96 to 1.96. Four items have a t value above 1.96 and six items have a value under -1.96. This is contradicted by the evidence from the Confidence Interval, which shows that the following items fall outside the 95% CI: 13, 20, 21, 25, 32, 34, and 35. As the CI is the more accurate statistic (Wu & Adams, 2013), we accept this result. The decision rule is to reject items only if they show misfit in BOTH statistical significance (CI), and effect size (MNSQ). Hence, we retain all the items except for 35.

Table 5.2
Item Fit T Statistics

Item	MNSQ	Confidence Interval		T
Item 1 ('ComprehensionOfProb-1')	1.05	0.93	1.07	1.5
Item 2 ('MappingToQuant-1')	1	0.95	1.05	0
Item 3 ('Strategy-1')	1.02	0.83	1.17	0.2
Item 4 ('Calculation-1')	1.02	0.85	1.15	0.2
Item 5 ('SoluVeri&Com-1')	0.98	0.94	1.06	-0.8
Item 7 ('ComprehensionOfProb-2')	0.94	0.93	1.07	-1.6
Item 8 ('MappingToQuant-2')	0.99	0.94	1.06	-0.2
Item 9 ('Strategy-2')	0.97	0.89	1.11	-0.5
Item 10 ('Calculation-2')	0.97	0.91	1.09	-0.6
Item 11 ('SoluVeri&Com-2')	0.92	0.9	1.1	-1.6
Item 13 ('ComprehensionOfProb-3')	1.13	0.95	1.05	4.6

Item 14 ('MappingToQuant-3')	0.96	0.93	1.07	-1.1
Item 15 ('Strategy-3')	1.01	0.9	1.1	0.2
Item 16 ('Calculation-3')	0.97	0.93	1.07	-1
Item 17 ('SoluVeri&Com-3')	0.94	0.94	1.06	-2.2
Item 19 ('ComprehensionOfProb-4')	0.98	0.94	1.06	-0.6
Item 20 ('MappingToQuant-4')	1.1	0.94	1.06	3.2
Item 21 ('Strategy-4')	0.93	0.95	1.05	-2.4
Item 22 ('Calculation-4')	1.03	0.92	1.08	0.8
Item 23 ('SoluVeri&Com-4')	1.14	0.85	1.15	1.8
Item 25 ('ComprehensionOfProb-5')	0.93	0.95	1.05	-2.5
Item 26 ('MappingToQuant-5')	1.04	0.94	1.06	1.5
Item 27 ('Strategy-5')	0.97	0.93	1.07	-0.9
Item 28 ('Calculation-5')	0.88	0.88	1.12	-2
Item 29 ('SoluVeri&Com-5')	1	0.94	1.06	0
Item 31 ('ComprehensionOfProb-6')	1.09	0.81	1.19	1
Item 32 ('MappingToQuant-6')	0.92	0.94	1.06	-2.8
Item 33 ('Strategy-6')	0.9	0.88	1.12	-1.8
Item 34 ('Calculation-6')	0.9	0.95	1.05	-4.1
Item 35 ('SoluVeri&Com-6')	1.25	0.9	1.1	4.4
Item 37 ('SoluVeri&Com-7')	0.97	0.95	1.05	-1
Item 38 ('SoluVeri&Com-8')	1.06	0.94	1.06	2

5.1.4 Response fit

Students sometimes get distracted while taking a test or for some other reason and do not answer the items as intended. This might result in some unusual result patterns and has a possibility of adversely affecting the model fit. This is especially a concern when administering a test that has lots of multiple-choice items as it is easy for students to get distracted or to randomly select an answer for some of the items.

Weighted Likelihood Estimate (WLE) data help us to identify respondents who had

unusual response patterns in a test. Note that, in the context of respondent fit, “overfit” is not usually considered a problem (Wilson, 2005). Table 5.3 shows the fit statistics for the respondents whose responses showed underfit above 1.96. This is $28/808 = 3.5\%$, which is quite close to the value of 5% we should expect if the persons are all fitting the Rasch model and we randomly generated the data. Thus, overall, respondent fit is very reasonable. Nevertheless, it is helpful to look at individual student responses to better understand their response pattern.

Table 5.3
Person Fit Statistics for Respondents Showing Underfit (Fit Statistics 2.00 and Above)

Person ID	Raw Score	Estimate	Standard Error	FIT
277	6	-1.84	0.49	3.50
179	9	-1.22	0.44	2.88
329	27	2.11	0.53	2.80
393	27	2.11	0.53	2.74
604	9	-1.22	0.44	2.63
28	26	1.85	0.50	2.63
455	9	-1.22	0.44	2.58
1	24	1.40	0.46	2.56
760	23	1.19	0.45	2.41
282	11	-0.85	0.42	2.37
267	11	-0.85	0.42	2.24
788	14	-0.35	0.41	2.22
189	24	1.40	0.46	2.18
400	25	1.61	0.48	2.17
804	10	-1.03	0.43	2.16
178	24	1.40	0.46	2.12
792	12	-0.68	0.42	2.10
191	14	-0.35	0.41	2.10

215	7	-1.62	0.47	2.06
498	9	-1.22	0.44	2.04
571	11	-0.85	0.42	2.04
151	26	1.85	0.50	2.04
180	26	1.85	0.50	2.04
148	21	0.82	0.43	2.02
746	25	1.61	0.48	2.02
343	16	-0.02	0.41	2.01
131	7	-1.62	0.47	2.00

Respondent number 277, whose student research ID (SRID) is 54114, shows the most underfit for the model. Table 5.4 shows the response pattern for this student in a revised item order where the items have been sequenced based on the revised item difficulty attained through the empirical data here.

Table 5.4
Response Pattern for Respondent 277 (SRID 544114) when Item Sequence Is Revised Based on Empirically Obtained Item Difficulty

Item	3	4	9	15	11	10	1	7	27	5	19	17
Response	0	1	0	0	0	0	0	0	0	0	1	0
Item	21	2	25	34	37	13	26	38	29	32	20	8
Response	0	0	0	0	0	1	0	0	0	0	0	1
Item	16	14	22	35	33	28	23	31				
Response	1	0	0	0	0	0	1	0				

The response pattern shows that the respondent has mostly given incorrect answers (represented by zeros) with only a few correct answers (represented by ones) randomly dispersed among the easy, medium and hard difficulty items. The correct answers do not have any pattern of distribution, so the student might have randomly guessed all the answers and got some of them right.

Table 5.5

Response Pattern for Respondent 179 (SRID 35057) when Item Sequence Is Revised Based on Empirically Obtained Item Difficulty

Item	3	4	9	15	11	10	1	7	27	5	19	17
Response	1	0	1	0	0	0	0	0	0	0	0	0
Item	21	2	25	34	37	13	26	38	29	32	20	8
Response	0	0	0	0	0	1	1	1	1	0	0	1
Item	16	14	22	35	33	28	23	31				
Response	0	0	0	1	0	0	1	0				

Table 5.5 shows the response pattern for the respondent 179 (SRID-35057) who had the second highest fit statistics, that is, underfit.

The response pattern shows that this respondent was only able to answer two of the easy items correctly, whereas s/he has correctly answered a number of the medium difficulty and hard items. S/he might have been distracted or inattentive for some reason at the beginning of the test and answered the items incorrectly. For the most difficult items, s/he only has a couple of correct answers which might be random guesses.

Table 5.6

Response Pattern for Respondent 329 (SRID 55036) when Item Sequence Is Revised Based on Empirically Obtained Item Difficulty

Item	3	4	9	15	11	10	1	7	27	5	19	17
Response	1	0	1	1	1	1	1	1	1	1	1	1
Item	21	2	25	34	37	13	26	38	29	32	20	8
Response	1	1	1	1	1	1	1	1	0	1	1	1
Item	16	14	22	35	33	28	23	31				
Response	1	0	1	0	1	1	1	0				

Table 5.6 shows the response patterns for the next respondent, PID 329 (SRID-55036).

The response pattern shows that this student has answered most of the items correctly with only a few random mistakes. This might be due to a lapse of concentration.

While such response patterns give us an overall idea about items for which the student's response does not match expectation, it is sometimes hard to comprehend which item responses are particularly out of the ordinary. For such cases, a 'Kidmap' – which is a modified version of the Wright map representing responses – can be used. Expressing test responses through a Kidmap gives us an opportunity to check which item responses are surprising. Figure 5.4 presents the Kidmap for respondent 329 (SRID 55036). The right-hand side of the map shows all the items the student has correctly answered, and the left-hand side shows those that were incorrect. The XXX on the map represents the student's location and the dotted lines near the XXX indicate a point beyond which the probabilities indicate surprise, hence they are called surprise lines (Wilson, 2005). In Figure 5.4, most of the student's responses should be on the lower right hand and upper left-hand quadrants. The unexpected responses in this Kidmap are on the left bottom quadrant with four items below the surprise line. It should be mentioned that the distance above or below the surprise lines indicate the relative improbability of a response (Wilson, 2005). Although items 35, 14, 29 and 4 all display unexpected responses for this student, item 4 is the furthest from the surprise line, which means this was the most improbable response from the student. It is likely that this particular response, along with the other three, had largely contributed to the underfit of this student's response to the model.

Case 329: 329 % Score: 84.38 Location(WLE): 2.107

=====Build: Jan 15 2019=====

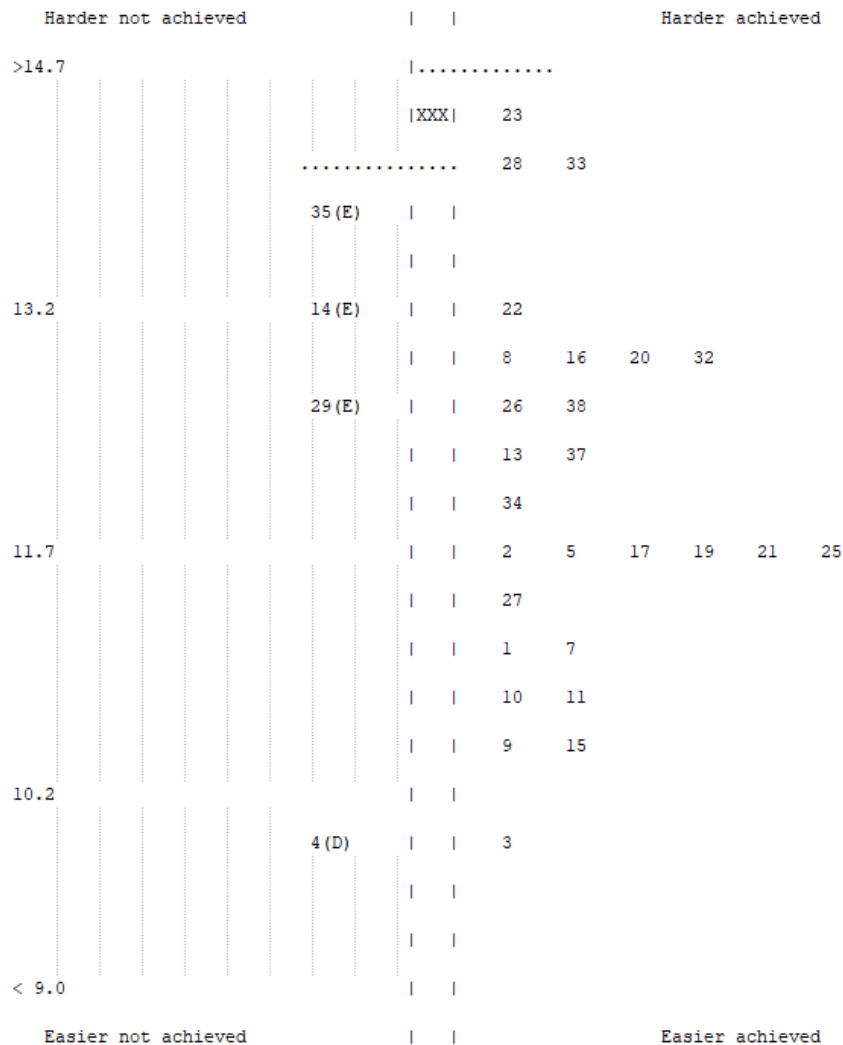


Figure 5.4. Kidmap for respondent 329 (SRID-55036)

Response patterns for the other respondents who had shown underfit (as detailed in Table 5.3) are similar to the ones discussed, they had a few zeros randomly placed in a string of ones or the other way around. So, most of these underfits could be attributed to lapses of concentration or randomly guessing correct answers.

5.1.5 Item-total correlation

Item-rest correlation is used to determine how individual items have contributed to the overall test. Items having positive item-rest correlations are suitable for the test whereas items with low or negative item-rest correlations need to be reviewed. Figure 5.5 shows the item-rest correlations for all the items. All the items have a positive item-rest correlation value except for item 35.

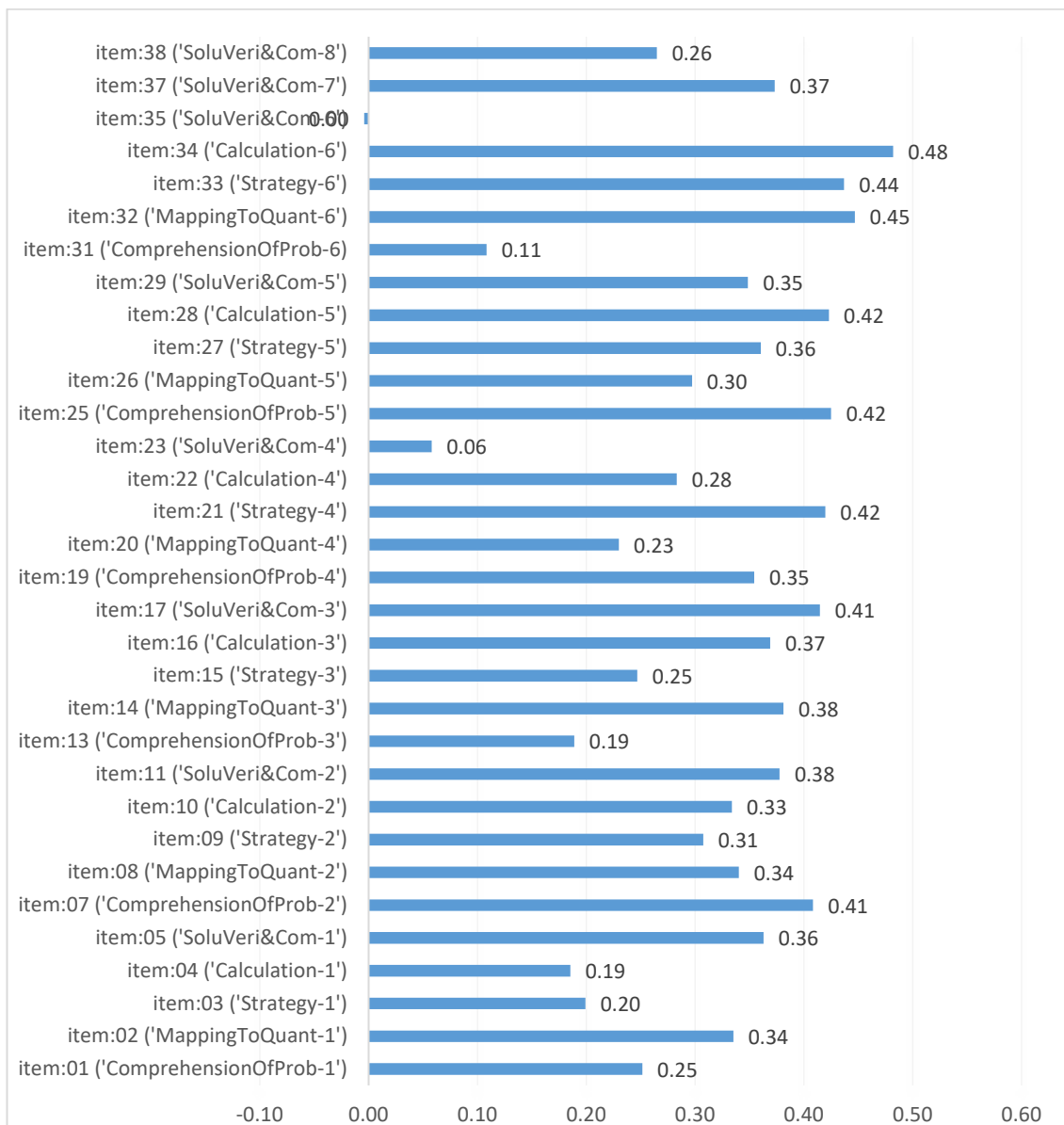


Figure 5.5. : Item-rest correlations of items in the final test

Most of the item-rest correlations are above 0.20. There are five items that have an item-rest correlation value lower than 0.20. These are Item 35 (0.00), Item 23 (0.06), Item 31 (0.11), Item 04 (0.19), and Item 13 (0.19).

Item 35 had an item-rest correlation value of zero. This item is not contributing to the reliability of the whole test and it might be advisable to delete or revise it in future iterations of the test. It is interesting that this item also had the worst underfit (Figure 5.3). Hence, there is even more evidence that we should delete this item.

Item 35: The headteacher of a school is planning a victory day celebration in her school and wanted each student to have a small Bangladeshi flag. These flags come in packets of 200s. The school has 656 students in total but 126 of them will not be present. The headteacher calculated there will be 530 students in the celebration and the school would not need to buy new flags. Which of the following alternatives show the correct calculation and result for the number of students and number of packets of flags required?

- a. $656 + 126 = 530$; 4 packets of flag
- b. $656 - 126 = 530$; 4 packets of flag
- c. $656 - 126 = 530$; 3 packets of flag (answer)
- d. $656 - 126 = 530$; 2 packets of flag
- e. None of the above

This item had a zero item-rest correlation in the pilot test as well and it was revised.

However, even the revised item has a very low value of item-rest correlation. Figure 5.6 shows the item-rest correlations for all the items if item 35 is deleted.

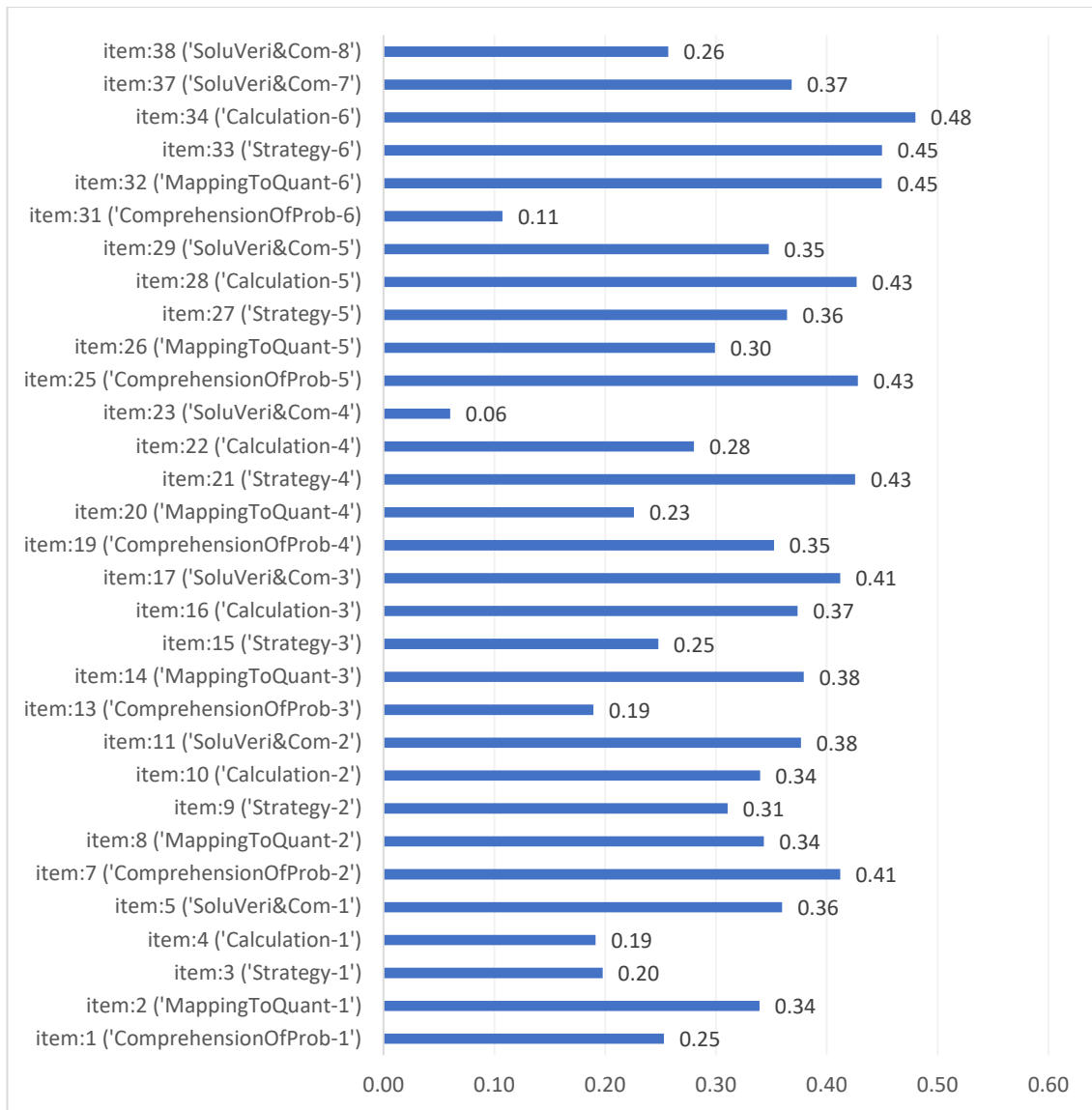


Figure 5.6. Item-rest correlations of items in the final test when item 35 is deleted

It is noticeable that the item-rest correlation for most of the items remains unchanged. The only items for which item-rest correlation value increases are items 10, 21, 25, 28, and 33. For all these items, the item-rest correlation value increases by 0.01. Such small amount of change does not impact the rest of the analysis for this study, but item 35 should be removed or revised for future iteration of administering this test as it is not contributing to the overall test.

Item 23 also had a low item-rest correlation value (0.06). It would probably be good to revisit this item in future iterations of this test.

5.2 Model fit: Unidimensional vs Multidimensional

While generally Arithmetic Word Problem Solving (AWPS) is considered as a unidimensional construct, this research hypothesised that it is a multidimensional construct that has five separate strands. A model fit comparison was conducted to check whether the unidimensional or multidimensional model is a better fit for the empirical data. Two models can be evaluated based on “changes in model deviance and number of parameters estimated” (Osteen, 2010, p 77). A chi-square statistic is calculated as the difference in deviance between two models with degrees of freedom (df) equal to the difference in number of parameters for the models. In such analyses, when statistically significant difference is found in model fit, the model with the smallest deviance is considered to be better (Osteen, 2010; Rabe-Hesketh & Skrondal, 2012; Wu & Adams, 2006).

To determine the best model fit for the empirical data collected in this research, both unidimensional and multidimensional model fit were analysed using *ConQuest*. Table 5.7 shows the model fit data for both models. The difference in deviance between the two models is 162.33 and the difference in number of estimated parameters is 14. Using the difference in parameters as *df*, the difference in deviance was determined to be statistically significant ($p < 0.0005$). There is a subtlety to determining the p value here. The extended Chi-square table has a critical value for up to $p = 0.001$. For $df = 14$ the critical value at $p=0.001$ is 36.123, which is lower than

the difference in deviance. However, according to Rabe-Hesketh and Skrondal (2012), while testing variance components, test statistic is conservative since they have a lower boundary of zero. As a result, they should be considered as one-sided and the p value reported should be divided by two. That is why the difference in deviance reported here is described to be significant at $p < 0.0005$.

Table 5.7

Comparison of Fit between Unidimensional and Multidimensional Model

Model	Deviance	Number of parameters	Change in Deviance	Degrees of Freedom (df)	p (significance)
5-Dimension	27914.19	47			
1-Dimension	28076.52	33	162.33	14	$p < 0.0005$

This shows that the model with the smaller deviance, that is, the multidimensional model (with five dimensions) represented the structure of the item response data more appropriately. From this it can be inferred that the AWPS construct is multidimensional. However, it is important to determine the effect size of this model comparison and the most directly applicable effect size here is the size of the correlations that are estimated in the multidimensional model calibration. The correlations among the different dimensions (shown in Table 5.8) range from 0.66 to 0.92.

Table 5.8

Estimated Correlations among the Dimensions in the Multidimensional Model

Dimensions	Dim 1	Dim 2	Dim 3	Dim 4
Dim 1 - Comprehension of the problem				
Dim 2 – Mapping conceptual relations to numerical representation	0.92			
Dim 3 – Selecting an appropriate strategy for solution	0.66	0.73		
Dim 4 – Performing the calculations	0.74	0.72	0.86	
Dim 5 – Verifying the solution and communicating the results	0.78	0.84	0.72	0.80

To put these values into perspective, let us consider the correlations between the different domains of the assessments conducted as part of the OECD PISA project 2018 (OECD, 2020). The correlation between Reading and Science ranged from 0.78 to 0.92 for different countries in the PISA 2018. The average correlation between Mathematics and Reading was 0.80, between Mathematics and Science was also 0.80, and between Reading and Science was 0.85. Taking these correlation values as a crude reference, the correlation between the different dimensions of the AWPS test presents a reasonable picture. That is, the high correlations are expected and do not undermine the finding that AWPS can be considered as a multidimensional construct.

5.3 Analysis of test reliability

The WLE person separation reliability calculated through *ConQuest* was 0.81, which indicates that the test was quite sensitive in distinguishing between high and low performing students. The EAP/PV reliability calculated for the test was 0.82. These values indicate the test has a high reliability.

In addition, Cronbach's Alpha was calculated—this is the raw score equivalent of the reliabilities shown above. Calculation of Alpha in *ConQuest* for this test returned a coefficient alpha value of 0.82.

To consider the multidimensional model where there are five separate dimensions within the test, the WLE person separation reliabilities and the EAP/PV reliabilities for each of the dimensions were calculated using *ConQuest*. Table 5.9 shows the WLE person separation reliabilities and EAP/PV reliabilities for the different dimensions of the multidimensional model. The EAP/PV reliabilities for the dimensions are good (≥ 0.70). However, the WLE person separation reliabilities range from 0.28 to 0.47 which show low to moderate sensitivity in distinguishing students of different ability. Nonetheless, while considering these outcomes, we need to keep in mind that almost all these dimensions had six items each (except Dimension 5, which had eight items). It is likely that person separation reliabilities for these dimensions will considerably increase if more items were added for assessing each dimension.

Table 5.9
Reliability Coefficients for the Different Dimensions of the Multidimensional Model

Dimension	WLE Person Separation Reliability	EAP/PV Reliability
D 1 – Comprehension of the problem	0.34	0.70
D 2 – Mapping conceptual relations to a numerical representation	0.28	0.73
D 3 – Selecting an appropriate strategy for solution	0.42	0.72
D 4 – Performing calculations on the numbers	0.47	0.74
D 5 – Verifying the solution and communicating the results	0.41	0.72

Person Standard errors

Determining person standard errors provides us with valuable insights about a test's reliability. This is calculated as the Standard Error of Measurement (SEM) of the estimate of the respondents' location (Wilson, 2005). Students' ability estimates (WLE) and the standard error of measurement were calculated in *ConQuest*. Figure 5.7 demonstrates how the standard errors were distributed for the different ability estimates. It shows that the standard errors ranged from approximately 0.4 to 0.625. This suggests that the errors in estimates of location are low and the test could be considered as reliable.

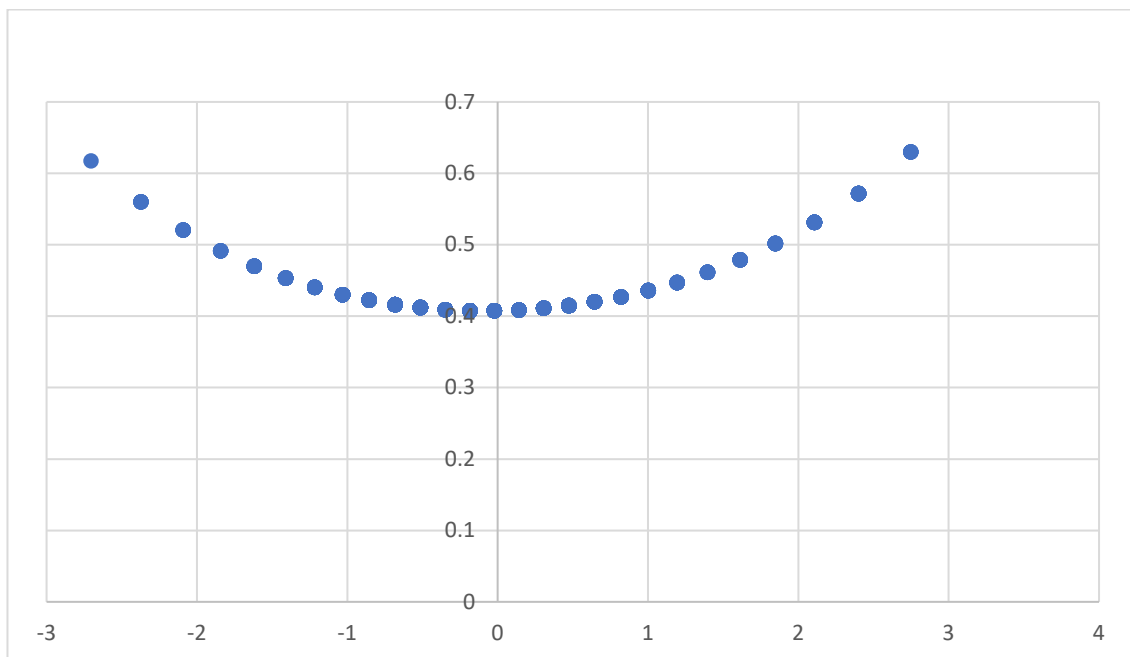


Figure 5.7. Standard Error of Measurement for the Final AWPS Test

5.4 Analysis of validity evidence based on relations with other variables

Two types of data were collected to gather evidence that can be compared with the AWPS test items' score: textbook items included in the test booklet and students' school test scores in Mathematics and Bangla. It was expected that the students'

performance in the AWPS test developed for this study should be positively correlated with their performance in the textbook items and their previous performances in Mathematics and Bangla tests in school. The textbook items assessed their performance in arithmetic word problem solving without separating the different hypothesised strands of AWPS, whereas the mathematics test scores in school examination were considered to be a proxy of their mathematical ability. In addition, comparability with the Bangla test scores was explored since two of the strands in the hypothesised AWPS framework (Comprehension of the Problem and Verifying the Solution and Communicating the Results) were considered to be related to reading and writing abilities. However, we need to keep in mind that it was not possible to review the items for the Mathematics and Bangla tests conducted by the schools. So, the appropriateness of these tests in measuring the students' mathematical or reading-writing ability is unknown.

Students' performance in the AWPS test (32 items) obtained from the Rasch analysis in logit values were converted to a scaled score using a linear transformation (where Scaled Score = Ability estimate in logit $\times$ 100 + 500). This transformation ensured that all the scores were positive and were to fall within a range of 200 to 800. This is helpful in communicating the results with the participating schools as they are not familiar with Rasch ability estimates and are unlikely to know how to interpret logit values (communicating the test results with the schools was one of the dissemination commitments of the research). These scaled scores were used for comparison of AWPS test performance and textbook item score. Similarly, the scaled scores were used for comparison with Bangla and Mathematics test scores from the schools.

In addition to comparing the scaled scores from the AWPS test, students' ability estimates from the different strands of the multidimensional model were compared with their textbook item score, Mathematics test score from schools and Bangla test score from schools. It should be mentioned that for these comparisons, the ability estimates for the separate strands were not converted to scaled score (since these are not going to be communicated with the schools). However, using these estimates does not affect the output as these were pairwise correlations.

5.4.1 Comparison of AWPS test and textbook item score

There were six items in the test booklet that were chosen from Mathematics textbooks of grades 4 and 5 (Bangladeshi textbooks). These items required addition and subtraction operations and presented the problems in a familiar format for the students. That is, unlike the items developed for this study, these were full arithmetic word problems that students are familiar with. Similar to most of the AWPS test items, the textbook items were presented as multiple-choice questions.

Using SPSS, correlation coefficients were calculated for students' total score in the six textbook items and their scaled score in the AWPS test (32 items).

Table 5.10

Correlation between Students' Scaled Score in AWPS Test and Textbook Item Scores

Grade	Number of Students	Pearson Correlation Coefficient
Both grades	808	0.62**
Grade Four	412	0.55**
Grade Five	396	0.65**

Note. ** Correlation is significant at the 0.01 level (2-tailed)

Table 5.10 displays the comparison of students' scaled scores in the AWPS test and their total score in the textbook items. As the Pearson correlation coefficients suggest, there is a positive correlation between students' performance in the AWPS test and their performance in the textbook items. In addition to calculating the correlation coefficient for the whole group of students, correlation coefficients were calculated for grade 4 and grade 5 students separately. Table 5.10 shows that the correlations for separate grades were also positive and high. Furthermore, all three correlation coefficients were statistically significant at the 0.01 level (2-tailed). Comparisons of students' ability estimates in the different dimensions of the Multidimensional model and their total score in the textbook items are presented in Table 5.11. These correlation coefficients were calculated in SPSS. If we consider students from both grades 4 and 5 together, the Pearson correlation coefficients range from 0.39 to 0.51. This is reasonably good. If we consider the grades separately, all the correlation coefficients are positive and within a reasonably good range (0.31 to 0.55). The analysis in SPSS showed all these correlations to be statistically significant at the 0.01 level (2-tailed).

Table 5.11
Correlations between Students' Textbook Item Scores and WLE estimates for the Dimensions of the Multidimensional Model

Grade	No of Students	Pearson Correlation Coefficient between total score in the textbook items and WLE estimates for the different dimensions				
		D1	D2	D3	D4	D5
Both grades	808	0.46**	0.50**	0.51**	0.47**	0.39**
Grade Four	412	0.41**	0.39**	0.42**	0.36**	0.31**
Grade Five	396	0.45**	0.55**	0.53**	0.51**	0.41**

Note. ** Correlation is significant at the 0.01 level (2-tailed)

5.4.2 Comparison of AWPS test and school test scores

Students' scaled scores in the AWPS test were compared with their school examination scores in Bangla (language) and Mathematics using SPSS. All the schools in this research were from the same district in Bangladesh and they use the same question papers for the school level examinations (i.e., the term examinations). It should be mentioned that for some students, Mathematics score in the school examination was not available due to reasons such as changing schools or absence during the examination. Among the 808 students taking the AWPS test, school Mathematics scores were available for 776 students (394 grade 4 and 382 grade 5 students). While calculating the Pearson correlation coefficients in SPSS, students without Mathematics test scores were excluded pairwise.

Table 5.12

Correlation Between Students' Scaled Score in the AWPS Test and their Mathematics Test Score in School Examination

Grade	Number of Students	Pearson Correlation coefficient
Both grades	776	0.40**
Grade Four	394	0.32**
Grade Five	382	0.56**

Note. ** Correlation is significant at the 0.01 level (2-tailed)

Table 5.12 displays the comparison of students' scaled scores in the AWPS test and their scores in school Mathematics tests. For this comparison, it is important to remember that grade 4 and grade 5 students had separate mathematics tests in their schools. The Pearson correlation coefficient for both grades shows a positive correlation between students' AWPS test performance and their school Mathematics test scores. The overall correlation coefficient (where all the students are considered

as one group) is also positive. Furthermore, these correlation coefficients were found to be statistically significant at the 0.01 level (2-tailed).

Table 5.13

Correlations between Students' Mathematics Test Score in School Examination and their WLE estimates for the Dimensions of the Multidimensional Model

Grade	No of Students	Pearson Correlation Coefficient between Mathematics test score and WLE estimates for the different dimensions				
		D1	D2	D3	D4	D5
Both grades	776	0.31**	0.31**	0.32**	0.29**	0.24**
Grade Four	394	0.30**	0.25**	0.25**	0.15**	0.16**
Grade Five	382	0.37**	0.43**	0.43**	0.48**	0.36**

Note. ** Correlation is significant at the 0.01 level (2-tailed)

Table 5.13 presents the Pearson correlation coefficients (calculated in SPSS) for comparison of students' Mathematics score in school examination and their ability estimates for the different dimensions of the Multidimensional model of AWPS. All the correlation coefficients are positive, and most are reasonably good. However, the correlation coefficients for grade 4 students are low compared to the grade 5 students. Without knowing more about the mathematics test that was used in grade 4 in the schools, it is difficult to understand what the reason could be for this low correlation coefficient value.

Students' scaled scores in the AWPS test were also compared with their school test scores in Bangla (language). Similar to Mathematics, Bangla scores were not available for some students and while calculating the Pearson correlation coefficient, they were excluded pairwise in SPSS. School Bangla test scores were available for 775 students in total (392 in grade 4 and 383 in grade 5).

Table 5.14

Correlation Between Students' Scaled Score in the AWPS Test and their Bangla Test Score in School Examination

Grade	Number of Students	Pearson Correlation coefficient
Both grades	775	0.49**
Grade Four	392	0.45**
Grade Five	383	0.50**

Note. ** Correlation is significant at the 0.01 level (2-tailed)

Table 5.14 shows the results from the comparison of students' AWPS test performance and their Bangla test scores from the school. The school tests were separate for grade 4 and grade 5, and the comparisons have been presented first with all the students considered as one group and then for the two grades separated. As noticeable from Table 5.14, the correlations are positive for all three comparisons. Moreover, these correlations are statistically significant at the 0.01 level (2-tailed).

The Pearson correlation coefficients for comparison of the students' Bangla test scores in school examination and their ability estimates for the different dimensions of the Multidimensional model of AWPS are presented in Table 5.15. All the correlation coefficients are positive and reasonably good. In addition, the analysis in SPSS showed all these correlations to be statistically significant at the 0.01 level (2-tailed).

Table 5.15

Correlations between Students' Bangla Test Score in School Examination and their WLE estimates for the Dimensions of the Multidimensional Model

Grade	No of Students	Pearson Correlation Coefficient between Bangla test score and WLE estimates for the different dimensions				
		D1	D2	D3	D4	D5
Both grades	775	0.37**	0.39**	0.38**	0.33**	0.34**
Grade Four	392	0.38**	0.29**	0.35**	0.21**	0.30**
Grade Five	383	0.32**	0.44**	0.37**	0.39**	0.34**

Note. ** Correlation is significant at the 0.01 level (2-tailed)

5.5 Analysis of validity evidence based on test content: Expert opinionnaire

Two types of validity evidence based on test content were utilised in this research: test design and specification while designing the test and expert opinion on the test items. Subject matter experts and assessment experts can be valuable resources while developing a test. They can help in determining whether a test or an item is assessing what it targets to assess.

As mentioned earlier, expert opinions were collected as part of gathering validity evidence for the AWPS items that have been developed. While developing the items, paneling was done with mathematical education experts from the University of Melbourne. After developing the final draft of items and piloting the items in one school, the items were revised, and final data collection was conducted with the students. While designing the items it was felt that to gather validity evidence for the items, it would be ideal to collect opinions of experts who are familiar with arithmetic problem solving. Since the items were administered in Bangladesh, it was felt that opinion of Bangladeshi experts regarding the items would be best suitable. So, data were collected from 24 experts who work in the Mathematics subject area in Bangladesh. These experts included subject specialists from the National Academy of Primary Education (NAPE) in Bangladesh (n=2), instructors from Primary Teachers' Training Institutes (PTI) (n=6), primary school head teacher (n=3) and primary teachers (n=13).

All the information collected were entered in SPSS, an error check was conducted and, after data cleaning, descriptive analyses were conducted. The

following sections discuss the opinions expressed by the experts on each set of items.

The sections have been presented here based on the strands the items represent.

Strand 1 Comprehension of the Problem Items

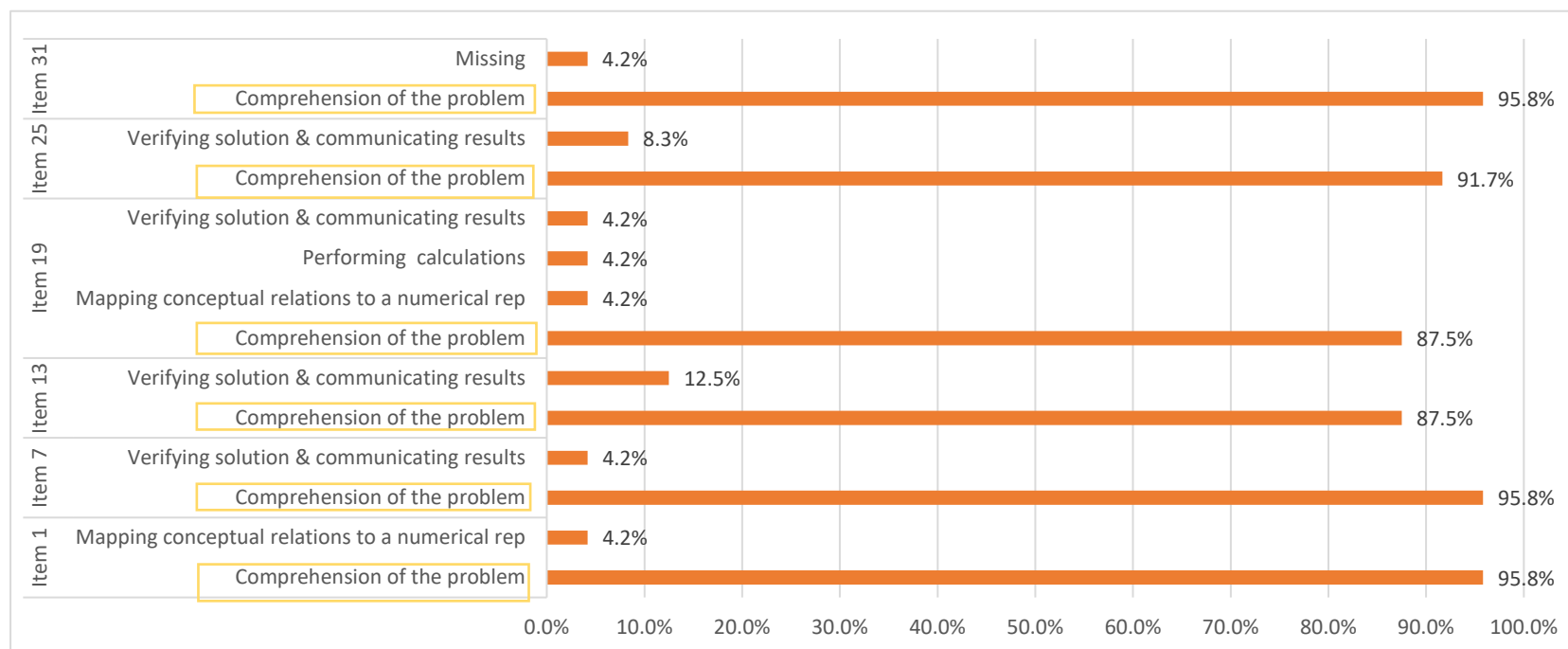


Figure 5.8. Expert opinion: Strand 1 – Comprehension of the Problem items

Figure 5.8 displays the experts' opinions about the 'Comprehension of the Problem' items.

For Items 1, 7 and 31, 95.8% of the respondents identified these as items belonging to Strand 1. Generally, for all the items these experts agreed about 90% of the time that they are Comprehension of the Problem items. Item 19 seems to be confusing the experts to some extent but still, 87.5% considered it to be an item from this strand.

Strand 2 - Mapping Conceptual Relations to a Numerical Representation Items

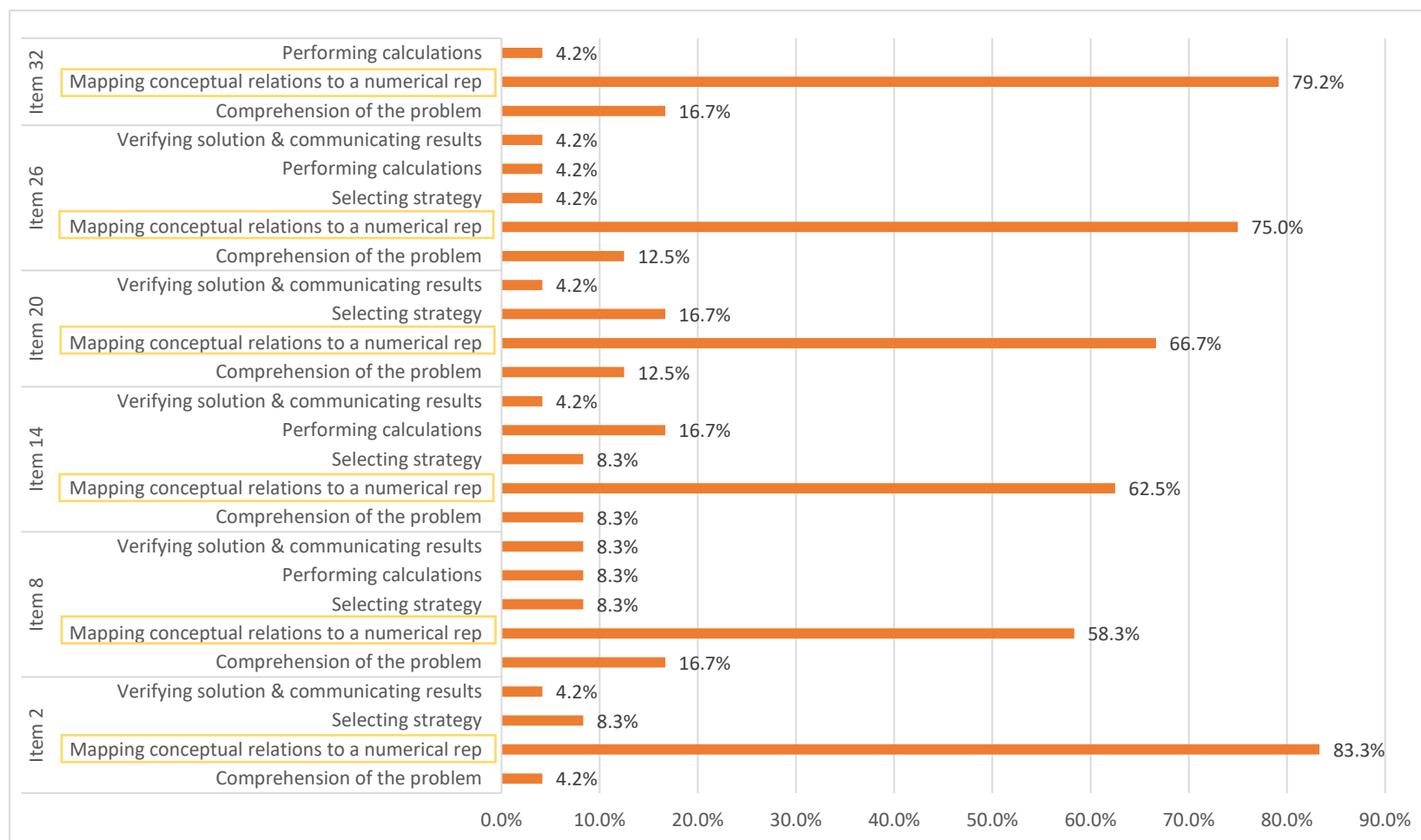


Figure 5.9. Expert opinion: Strand 2 – Mapping Conceptual Relations to a Numerical Representation items

Figure 5.9 shows the experts' opinion about the 'Mapping Conceptual Relations to a Numerical Representation' items. 75% or more experts identified items 2, 26, and 32 to be belonging to this area. Most of the experts agreed the other items also belonged to this area. Item 8 got the least number of experts (58.3%) agreeing that it was an item in this area. A number of experts considered this item (item 8) to be belonging to one of the other areas.

Strand 3 – Selecting an Appropriate Strategy Items

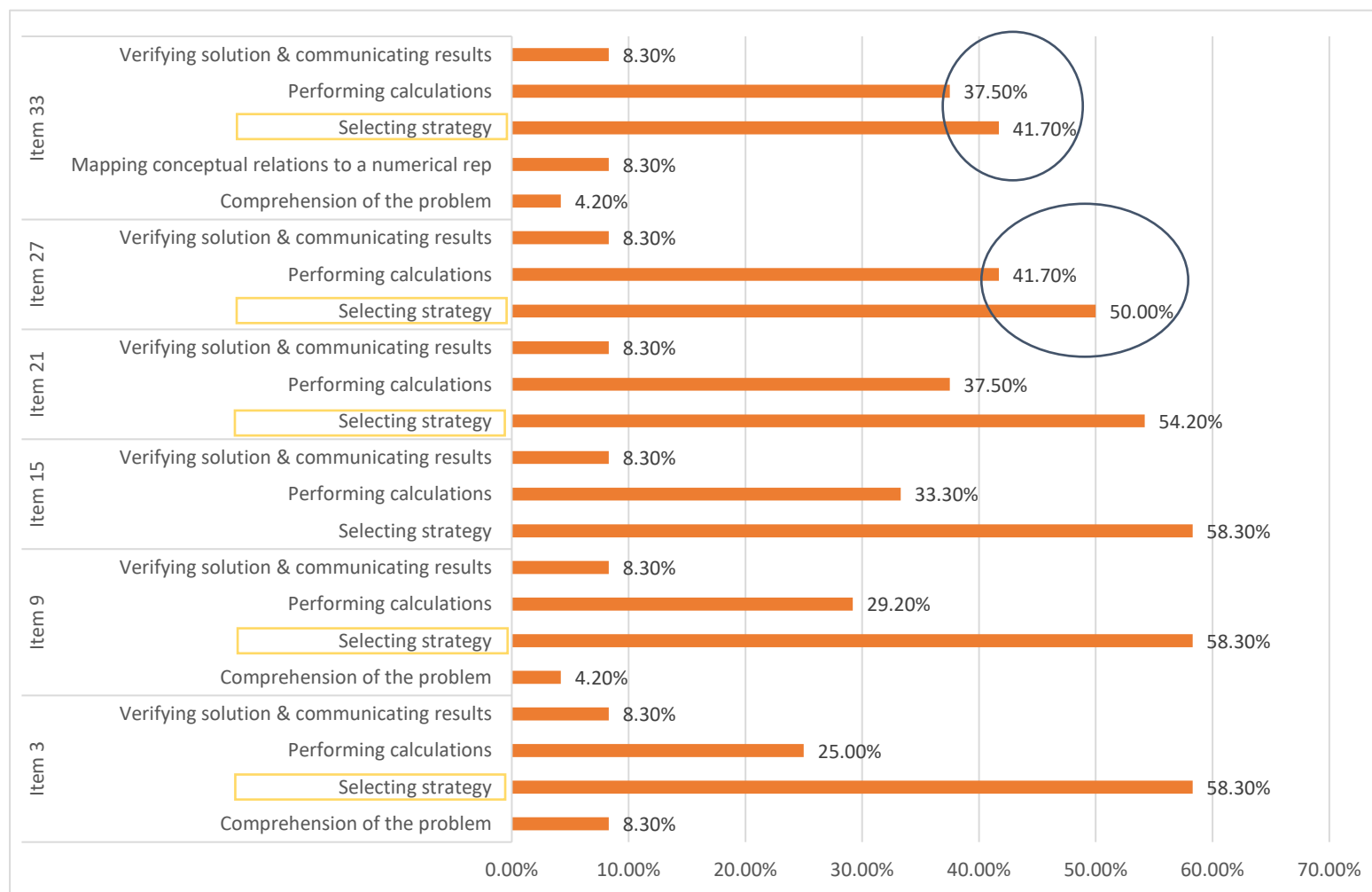


Figure 5.10. Expert opinion: Strand 3 – Selecting an Appropriate Strategy items

Figure 5.10 shows the experts' opinion about the 'Selecting an Appropriate Strategy' items.

For all the items, experts have predominantly agreed that they belong to this area.

However, it is noticeable how many of the experts identified these items as belonging to the 'Performing Calculations' area. For the first three items (3, 9 and 15) 58.30% of the experts agreed these are 'Selecting an Appropriate Strategy' items but there were 25%, 29.20% and 33.30% respectively considering items 3, 9 and 15 as 'Performing Calculations' items. For the last two items (27 and 33) the differences in the agreements of the experts were narrower. Half of the experts considered Item 27 as a 'Selecting an Appropriate Strategy' item whereas 41.70% thought it was a 'Performing Calculations' item. This gets even closer for item 33 where 41.70% of the experts thought it was a 'Selecting an Appropriate Strategy' item and 37.50% thought it was a 'Performing Calculations' item. These responses might be helpful in revising the items in future iterations.

Strand 4 – Performing Calculations Items

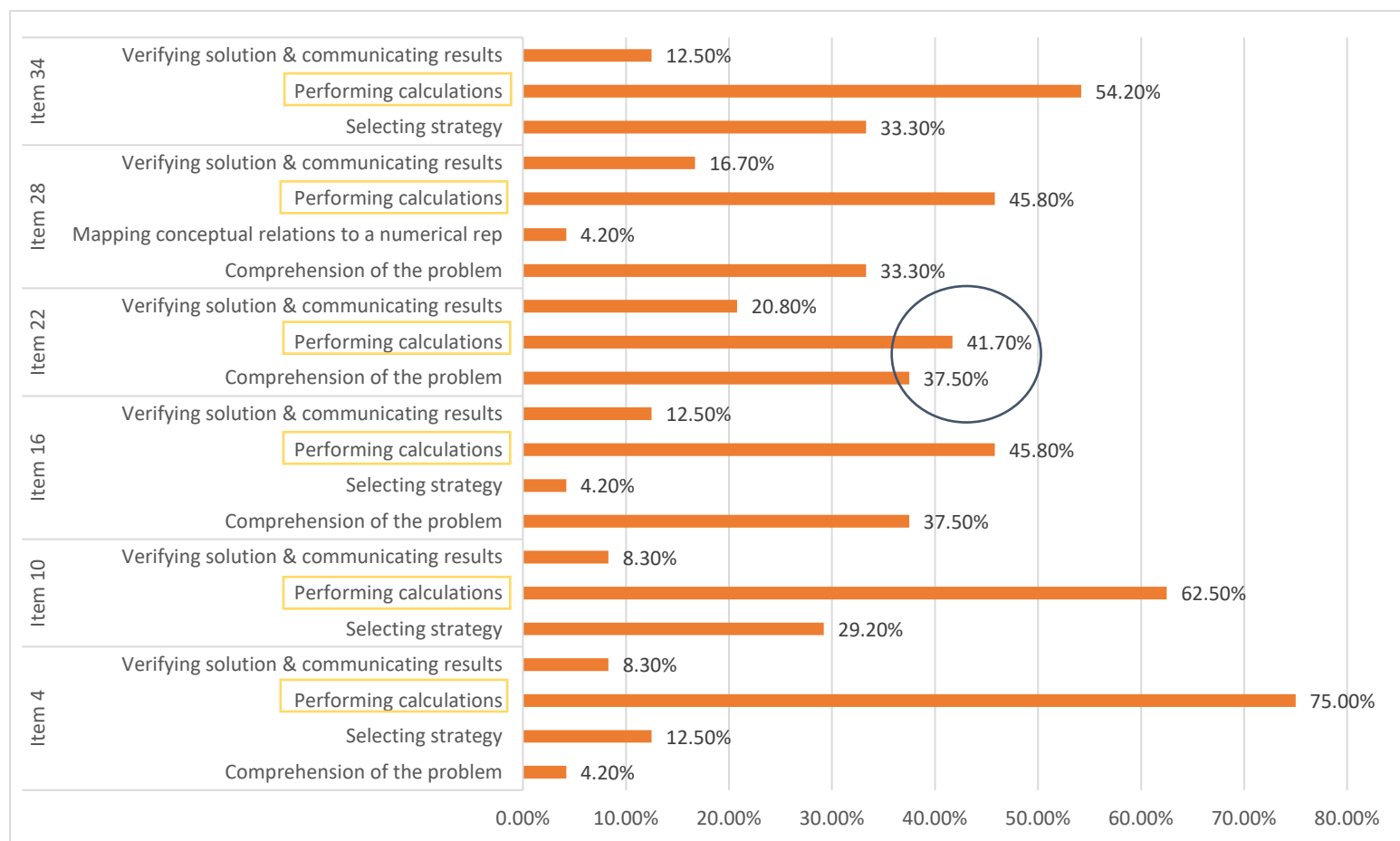


Figure 5.11. Expert opinion: Strand 4 – Performing Calculations items

Figure 5.11 depicts the experts' opinion regarding the 'Performing Calculations' items. All the items have been agreed to be belonging to this area by majority of the experts. Items 4, 10 and 34 are clearly identified as belonging to this area by experts with 75%, 62.5%, and 54.20% of them agreeing that 'Performing Calculations' is the area that these items are representing. The other items have a broader array of responses compared to these three. For item 16, 45.80% of the experts considered it as a 'Performing Calculations' item whereas 37.50% considered it to be a 'Comprehension of the Problem' item. Similarly, 41.70% of the experts agreed item 22 is a 'Performing Calculations' item but 37.50% chose 'Comprehension of the Problem' as the strand this item is representing. Item 28 also got 33.30% identifying it as a 'Comprehension of the Problem' item, though 45.80% agreed that it is a 'Performing Calculations' item. It might be worth more closely exploring why these items are being considered as representing another area in future administrations of this test.

Strand 5 - Verifying the Solution and Communicating the Results Items

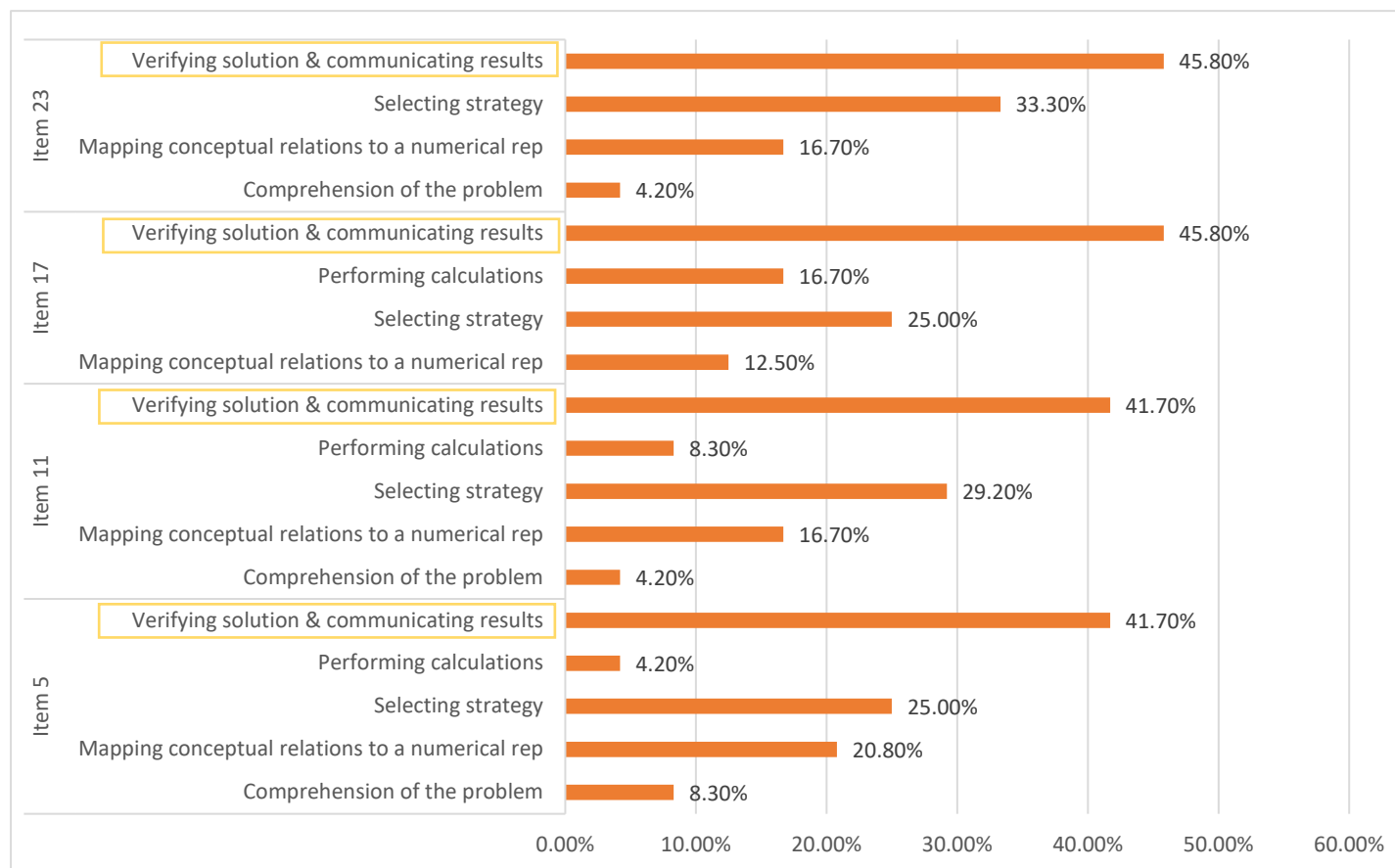


Figure 5.12. Expert opinion: Strand 5 – Verifying the Solution and Communicating the Results items (set 1)

Figure 5.12 shows the responses to the first four items of the 'Verifying the Solution and Communicating the Results' area. Although majority of the experts agreed that these items belong to this area, it is noticeable how many other areas have been chosen as relevant for these items. For all these items, more than 40% of the experts have agreed that they are 'Verifying the Solution and Communicating the Results' items. However, for the first two items (5 and 11), the experts seem to have brought up all the areas of AWPS in their responses. Similarly, for the other two items (17 and 23), at least three additional areas have been chosen by the experts. For all these items, 'Selecting an Appropriate Strategy' seems to be the second leading choice among them.

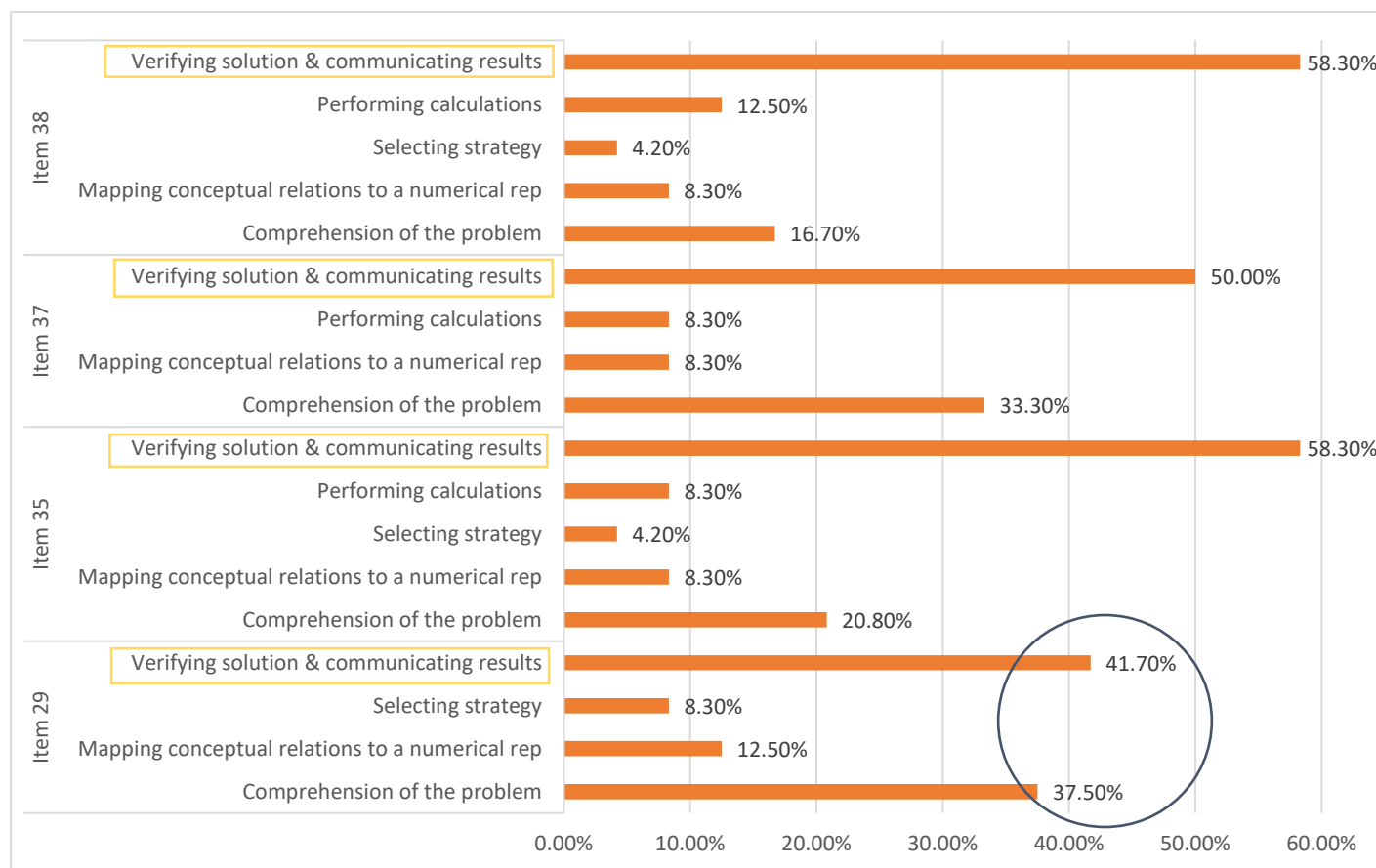


Figure 5.13. Expert opinion: Strand 5 – Verifying the Solution and Communicating the Results items (set 2)

Figure 5.13 shows the responses to the last four items of the 'Verifying the Solution and Communicating the Results' area. Again, a majority of the experts have agreed that these items belong to this area. Items 35, 37 and 38 have 50% or more experts identifying them as 'Verifying the Solution and Communicating the Results' items. For item 29, 41.70% of the experts identified it as a 'Verifying the Solution and Communicating the Results' item whereas 37.50% of them identified it as a 'Comprehension of the Problem' item. It is noticeable that for these items, responses show 'Comprehension of the Problem' as the second leading choice.

From the analysis of experts' responses to different items discussed above, it is apparent that generally they agreed to the hypothesised item strands. For all the items, the highest percentage of response was always in agreement with the strand the item was designed to represent. However, some of the items had a close second nomination for a strand that was different from what was hypothesised. For example, item 22 (Performing Calculations item, had Comprehension of the Problem as a close second), item 29 (Verifying the solution and Communicating the Results item, had Comprehension of the Problem as a close second), and item 33 (Selecting an Appropriate Strategy item, had Performing Calculations as a close second) all had a close second nomination from the experts. It is possible that a combination of skills (i.e., more than a single skill) was required to answer these items, though they were of varying prevalence. The experts who selected the strands other than what these items were designed for might have focused more on those skills. This is an area that could be further explored in future design of the test and data collection.

Another interesting point that can be noted from the data analysis is how the responses were quite sharply in line with the item design at the beginning (i.e., a large portion of experts choosing 'Comprehension of the Problem' for items designed to belong to that area) but gradually became more scattered for the later strands. This might be evidence that Arithmetic Word Problem Solving is a process and when the respondents reached the items in the later strands, all the earlier strands came into play to some extent.

It was also felt that the explanation of the theory that was attached with the expert opinionnaire might be too brief to clarify the different areas of Arithmetic Word Problem Solving. It was noticeable how some of the experts who had an in-depth discussion about the theory with the researcher, answered differently (i.e., more in line with the proposed design of items), compared to the experts who only had access to the brief explanation of the theory with the opinionnaire. Another issue might have been test fatigue. The respondents might have found it tiresome to identify the strands for all of the 32 items and might have randomly selected answers towards the end of the opinionnaire.

Chapter summary

This chapter discussed the analysis of quantitative data from the final phase of data collection. It included data analysis for the validity evidence based on the internal structure of the test (e.g., the Wright map, item fit, response fit, etc.). It also outlined the analysis of model fit (unidimensional vs multidimensional) and analysis of test reliability. In addition, analyses of validity evidence based on relation with other variables (e.g., comparison of AWPS test performance and textbook item scores,

comparison of AWPS test performance and school test scores in Mathematics and Bangla) were also discussed. The final section of the chapter depicted the analysis of one source of validity evidence based on test content, that is, the analysis of expert opinionnaire data.

Chapter 6: Analysis of qualitative data

As discussed previously, cognitive interviews were conducted with the respondents in this research to collect validity evidence based on their response processes. This chapter discusses the analysis of the qualitative data from the cognitive interviews collected in the final phase of the research. First, a brief overview of how the cognitive interviews were conducted and how the data were prepared for analysis are discussed. Then the data analysis is presented under separate headings for the different strands.

The quantitative data collection of this research included administering the AWPS test to grade four and five students at the selected schools. In addition to administering the test, cognitive interviews were conducted with a selected sample of students. The sample for the cognitive interviews in the final phase had 20 students selected from three schools through convenience sampling. The sample consisted of nine grade four students and 11 grade five students. Eleven of the respondents were female and nine were male. All the interviews were done after the students had already taken the test. Parents' consents were collected to include the students as respondents of the cognitive interviews, students' consents were also secured and then the interviews were done on the school premises following a standard operating procedure (discussed in Chapter 3). After completing the cognitive interviews, all responses were first collated in separate files (first in paper files, then electronic copies) for each student. To prepare the data for analysis, individual files were created for each item and students' responses to the items brought together. Information on the respondents' grade, their merit position in their classes and their gender were

retained in these collated files through codes. For example, each response had a code such as 34101_F attached to it. This code (34101_F) is based on the student's research ID. It shows that this student is from school 3 (research ID allocated to the school), grade 4, section 1; her merit position in her class is 01 and she is a female. This coding made it possible to refer to the students' responses in the interview as well as to the AWPS test (if necessary) and gave an indication of the students' merit positions while doing the qualitative analysis.

The focus of the cognitive interviews was to understand the students' response processes for the items in the AWPS test. Hence, the analysis focused on eliciting whether the response processes were similar to what was expected for each item, that is, whether the items made the students use the cognitive processes they were hypothesised to represent.

The analysis process described by Miller et al. (2014) (discussed in Chapter 3) was used for analysing the cognitive interview data. This analysis process has five steps: (i) conducting the interview, (ii) producing interview summaries, (iii) developing a thematic schema, (iv) developing an advance schema, and (v) making conclusions. As Miller et al. (2014) suggest, the first three steps of this analysis process often take place simultaneously. The analysis presented in this chapter outlines the common themes across responses for the same items (the third step). These themes are directly extracted from the cognitive interviews. Some quotes from the respondents (translated from Bangla) have also been included, while presenting the analysis where relevant. Then cross-group comparisons of the identified themes are discussed (the fourth step). Students with different abilities are identified as different groups for the

cross-group comparison. Their merit positions in the class (which is determined based on school examinations) are taken as a proxy of their abilities. Students with class roll 1 to 5 were considered to be high ability students, class roll 6 to 15 to be medium ability students and class roll 16 and beyond to be low ability students. This grouping was adopted based on discussions with the mathematics teachers of the sample schools. (More details about the groups are described in Appendix 6.1). Finally, conclusions are made about the item's performance based on the themes developed (the fifth step). The following sections of this chapter discuss the analysis of data for the items in different strands.

6.1 Strand 1 - Comprehension of the problem

There were six items in the test representing this area of AWPS. These items are: Item 1, Item 7, Item 13, Item 19, Item 25, and Item 31. Students' interview responses for these items will be discussed in this section.

Item 1: Jahid had 5 apples. Koli gave him 7 more apples. How many apples did Jahid have in the beginning?

- a. 12 b. 7 c. 5 d. 2 e. Not enough information to tell

Almost all the respondents were able to correctly answer Item 1. They could all read the information given here. Most students discerned the information provided in the stem and selected the correct answer saying things like "C is the correct answer because Zahid had 5 apples in the beginning" (34101_F) or "C is the correct answer because the item says so" (35003_M). However, some students answered the item incorrectly. They could read the stem without any difficulty but could not understand

what the item was asking. They added the two numbers (7 and 5) and selected 12 as their answer. Cross-group comparison of the themes showed that except for one respondent from fourth grade (94306_F), who was from the medium ability group, all the others who wrongly selected “a” as their answer were low ability students according to their merit position in class (e.g., 34245_M, 95224_F, and 95230_F).

Item 7: 3 of Raju’s friends gave him some coins. First, Samia gave him 45 coins; Jakaria gave him 80 coins; and then Lisa gave him 25 coins. How many coins did he get from the last one of his friends?)

- a. 125 b. 80 c. 45 d. 25 e. None of the above:

Except for two students all the students were quick in identifying the correct answer for item 7. They read the item without any difficulty and selected “d. 25” as their answer saying things like “the last friend Lisa had given Raju 25 coins” (34101_F) or “the last friend had given Raju 25 coins so d is the correct answer” (25031_M). It seems they were able to understand what the stem had asked for. Both the students who answered the item incorrectly were able to read the stem without any trouble. One of them (94366_F) chose “c”, that is, 45 as her answer and the other (95224_F) chose “b”, that is, 80 as her answer. Though they could read the item without any trouble, they had difficulty understanding the sequence of events described within the stem. Cross-group comparison suggested that both students incorrectly answering the items were from the low ability group.

Item 13: Jamal has more toys than Bakar who has more toys than Rimi but less than Rafi. Which of the following is true?

- a. Bakar has more toys than Rafi
- b. Rafi has more toys than Rimi
- c. Rimi has more toys than Rafi
- d. Rafi has more toys than Bakar
- e. None of the above

Item 13 seemed to have confused the students the most. It was designed as a medium difficulty item, but it had two correct answers. Selecting either of the correct options would be considered as a correct answer. However, only about half of the students (11) could answer this item correctly. Three themes emerged from analysing the interview data for this item. First, *full information retention and comprehension* displayed by the group of students who were able to keep all the information in their memory, could compare them and deduced the correct result (either “b” or “d”). Second, *partial retention and inability to comprehend* shown by the students who skipped some of the information given in the stem and selected “a” as their answer. About five students interpreted the information they read in the stem to mean that Bakar has more toys than Rafi and selected “a” as their answer. When asked why they thought this was the correct answer, they simply said because the stem said so (e.g., 34245_M). The third theme was *information overload and failure to comprehend* as for some respondents this item seemed to be a lot of information to process. One respondent (35006_F) got very confused and nervous. She kept reading the stem and the alternatives again and again and went silent trying to determine the answer. After a long time of silence, she was still unable to come up with an answer. So, the interviewer asked her to skip the item and move on to the next one. Comparing the

themes for this item's narratives across the different ability groups does not show any meaningful result. Both high ability and low ability students had difficulty in correctly answering this item. Respondents who skipped parts of the information and selected "a" as their answer included 34101_F, 34245_M, 35001_M, 25003_M, and 94306_F.

Item 19: A cricket bat weighs 2 kilograms, a pencil weighs 50 grams, a ruler weighs 100 grams, a television weighs 6 kilograms, and a tomato weighs 30 grams. A cricket bat has the highest length among these but which of these objects weighs more than the cricket bat?

- a. Pencil
- b. Ruler
- c. Tomato
- d. Television
- e. None of the above

The prevalent theme for item 19 is the students' *ability to compare between different units of measurement*. About a quarter of the students failed to answer this item correctly. Those who correctly answered the item were able to retain the information given in the stem and compare among the different units. After reading the stem they said things like "answer is d because the television weighs 6 kilograms" (35001_M) with emphasis on kilogram. All the students who incorrectly answered the item could read the item without any problem, but they could not comprehend the information provided in the stem as they were unable to understand the different units of weight measurement mentioned here. They selected "b" as their answer and while selecting, said things like "because it weighs 100 grams" (35006_F) (with emphasis on 100) or "because ruler has the largest number" (94306_F).

Cross-group comparison of the theme suggested that students from all three ability groups made the same mistakes for this item.

Item 25: Khulna is to the west from Dhaka, Narsingdi is to the east of Dhaka. Sylhet is further east from Dhaka. Which of these cities is the furthest towards the west?

- a. Sylhet
- b. Dhaka
- c. Narsingdi
- d. Khulna
- e. Not enough information to tell

Almost all of the students were able to correctly answer item 25. They were all able to read the item without any difficulty. The prominent theme from the analysis of response narratives for this item was students' *ability to understand directions from textual descriptions*. While responding to this item most students selected "d. Khulna" as their answer and said things like, "Khulna is the correct answer because the item mentions Khulna is to the west from Dhaka and all the other cities are to the east from Dhaka" (34101_F), or "Khulna is to the west from Dhaka and all the other cities are to the east from Dhaka" (34102_F) or simply "the answer should be Khulna" (25011_M). Only one student was unable to correctly answer the item. She read the stem without any trouble. But seemed to have difficulty in understanding the information. She read the stem and the alternatives a few times and then selected "e. None of the above" as her answer. She kept saying "I don't know, this is confusing..." (95230_F).

Cross-group comparisons suggest that the student who was unable to answer the item correctly was a lower ability student (95230_F).

Item 31: Water from a swimming pool has some salt in it. Tap water has the least amount of salt. But there is much salt in water from the Atlantic or the Pacific Ocean. Which water contains the most salt?

- a. Tap water
- b. Swimming pool water
- c. Water from the Atlantic
- d. Water from the Pacific
- e. Not enough information to tell

While answering Item 31, all students were able to read the stem and the alternatives without any trouble. However, only 8 of the 20 students interviewed could answer the item correctly. There were four themes that could be distinguished from analysing the responses: *considering all the available information including absence of information, forcing a choice based on available information, randomly choosing an answer without full comprehension, and struggling with the item altogether and failing to choose any answer.*

Respondents who correctly chose “e. Not enough information to tell” as their answer seemed to have comprehended the stem and the alternatives properly. They realised there was some information missing. Respondents who chose this as their answer said things like “water from both Atlantic and Pacific Oceans have lots of salt. As there is no way to differentiate between these two here, ‘e’ that is, ‘Not enough information to tell’ is the correct answer” (35001_M) or “water from both Atlantic and Pacific Oceans have lots of salt. So ‘e’ should be the answer” (25001_M) or “e is the

correct answer, because water from both Atlantic and Pacific Oceans have lots of salt but the item does not mention which one has more” (25011_M).

Responses in this item highlighted the fact that it was important for the students to focus both on the stem and the alternatives to correctly answer the items. Several students (9) chose either Atlantic or Pacific Ocean as their answer. They seemed to have understood the idea that these two oceans have lots of salts. However, they thought either of these two would be the correct response and they did not consider ‘not enough information to tell’ as a feasible answer option. When asked why they chose one of these as the answer, most of them could not explain. Some (e.g., 34107_M) just said, “because it has lots of salt and I saw it among the alternatives”. This group of students seemed to have partially comprehended the item.

Two students chose swimming pool as their answer. It is likely that they could not fully comprehend the item, got confused and chose swimming pool as it was mentioned first in the stem. When they were asked why they thought swimming pool was the correct answer, they could not explain. One of them (34102_F) said she chose it because she knew water from swimming pool has lots of salt.

For one student (94306_F) the item was simply too difficult to comprehend. She could read the stem and the alternatives but could not compare among the different options. She got very confused and spent a considerable amount of time reading and re-reading the item. She was getting uncomfortable, so the interviewer asked her to move to the next item.

Cross-group comparison for this item showed that all the students who correctly answered this item were high or medium ability students. However, there were some students who were from these two groups but could not comprehend the items correctly.

6.1.1 Summary of analysis for Strand 1 items

Overall, the response processes of the respondents perceived through the cognitive interviews demonstrated that most of the items in this strand are working as intended and assessing students' problem comprehension ability. However, Item 13 was confusing for some of the respondents and could be considered for revision for future iteration of the test. For the other items, the interviews showed that when students could answer an item correctly, it was because they could comprehend the information given in the item. When they were unable to answer correctly – even if they were able to read the item without any trouble – they were unable to discern the information given in the item. In addition, the items were more difficult for lower ability students than for higher ability students, as expected. The difficulty level of the items also loosely followed the hypothesised difficulty, that is, the first couple of items were easier and the later items were more difficult. However, item 25 seemed to be easier for the students than its hypothesised difficulty.

6.2 Strand 2 - Mapping conceptual relations to a numerical representation

There were six items in the test representing this area of AWPS. These items are: Item 2, Item 8, Item 14, Item 20, Item 26, and Item 32. Students' interview responses for these items will be discussed in this section.

Item 2: Razia has 15 candles. Daud has 7 candles less than Razia. Which of these number sentences is appropriate for calculating how many candles Daud has?

- a. $15 + 7$
- b. $7 + 15$
- c. $7 - 15$
- d. $15 - 7$
- e. None of the above

Item 2 was designed to be an easy item within this area and most of the students were able to answer it correctly. All of them were able to read the item without any trouble. There were two prominent themes within the students' responses. The first theme was *phrase triggered responses* and was represented by students focusing on the 'less than' part within the stem and correctly determining that this should be a subtraction. This group of students was able to identify the answer option that presented the smaller number being subtracted from the larger number. They chose "d" as their answer and said things like "d should be the answer because Daud has 7 candles less than Razia" (34107_M), or "d is the correct answer because Daud has 7 candles less than Razia, so it should be a subtraction" (35001_M). They were emphasising the 'less than' part. Another theme was students *focusing on the numbers from the stem*. Students whose responses fell in this theme (four of them) could read the item but answered it incorrectly selecting "a" (i.e., $15+7$) as their answer. They did not focus on the *less than* part within the stem and simply went for the first alternative that mentioned both the numbers from the stem. These students failed to correctly answer the item as they could not properly map the numbers with the relationship among them as per the stem.

Cross-group comparison for the themes suggested that some students from all ability groups were able to map the relation correctly. However, all the students mistakenly mapping the relation as an addition were either from the medium or low ability group students.

Item 8: Latif has 138 plants in his house. Mitu has 112 plants more than him. Which of these number sentences would be appropriate for calculating how many plants Mitu has?

- a. $138 - 12$
- b. $112 - 138$
- c. $140 - 12$
- d. $140 - 138$
- e. None of the above

Item 8 was designed as an easy item within this area and twelve of the twenty students interviewed were able to answer it correctly. Similar to the last discussed item (item 2) there were two themes extracted from the interview data. One theme (*phrase triggered responses*) represented students who focused on the phrase ‘more than’ mentioned in the item and determined this should be an addition. They searched for a number sentence among the alternatives that had addition and when they could not find one, they correctly selected ‘None of the above’ as their answer. While answering the item they were saying things like, “to determine how many plants Mitu has we need to add 112 with 138 because Mitu has 112 plants more (saying more with emphasis) than Latif. So, the correct answer should be $138+112$, but it is not in the alternatives. So ‘e’ ,that is, ‘None of the above’ should be the answer”

(35001_M) or, “because Mitu has 112 plants more (emphasised) than Latif, this should be an addition. So, the answer here should be e. none of the above” (34102_F).

The other theme, that is, *focusing on the numbers from the stem*, was represented by students who could read the stem but were unable to determine the correct mathematical relation between the numbers. These students seem to be exclusively focusing on the numbers mentioned in the stem and overlooking the relationship among the numbers. They focused on the alternative that had both the numbers from the stem and disregarded the relationship. All of them chose “b” (i.e., 112 – 138) as their answer and did not think aloud why this was the right answer. When probed about their selection, they said things like, “b should be the answer because it has both the numbers mentioned in the item” (35008_F).

This specific phenomenon might be related to the choice of the type of items (i.e., using MCQ items). When students are given alternatives in a multiple-choice item, some of them might be tempted to choose an alternative that has both the numbers from the item. However, in order to identify the correct answer, they should also be looking into the relationship mentioned in the stem to determine what the correct operation among the numbers ought to be.

Cross-group comparison of the themes showed that there were students from all ability groups who correctly answered the item. However, all the students exclusively focusing on the numbers from the stem and incorrectly determining the required operation from the stem were from the Medium or Low ability groups.

Item 14: Two hundred and twenty-three boys, one hundred and thirty-five girls, and twelve teachers participated in a school trip. Which of these number sentences is correct for finding out the total number of people going on the trip?

- a. $223 + 135$
- b. $135 + 12 - 223$
- c. $223 - 135 - 12$
- d. $12 + 135 + 223$
- e. None of the above

All students were able to read Item 14 without any difficulty. While analysing this item, four separate themes emerged. The first theme was again *phrase triggered responses*. For this item, some students focused on the word ‘total’ in the stem while looking for an answer. More than half of the students interviewed were able to answer this item correctly. For them, the word ‘total’ seemed to have triggered an instinct to look for a number sentence that had addition. Once they determined that the operation for this problem should be addition, they were able to identify “d” as the correct answer. Though “d” had the numbers in a different sequence compared to the sequence in the stem, these students were able to confidently identify it as the correct answer. While solving the item they said things like, “since this asked for the total (emphasising the total) number of people going for the school trip, this should be an addition” (35001_M). The next theme that was visible was students *focusing on the sequence of the numbers and ignoring the operation*. Five of the students used this approach and chose “c” as their answer. These students seemed to be preoccupied to look for the numbers in the exact sequence as mentioned in the stem and did not closely check what the relation among the numbers should be. Consequently, they jumped to an answer once they identified the numbers from the stem. When asked

why “c” was the correct answer, the answer was “because it has the correct numbers” (94306_F).

The third theme was a combination of the previous two. Students whose responses were categorised in this theme were *focusing on both operation and the sequence of numbers*. However, respondents in this group ignored the third number from the item. They understood this should be an addition, but they also could not imagine the numbers out of sequence from how they were presented in the stem. So, they chose “a” as their answer which was an addition and had the first two numbers from the stem in sequence. Three of the respondents chose “a” as their answer. The final theme represents students *randomly choosing an answer* and being unable to explain why they chose it. One student (35006_F) chose “b” as her answer which did not have the correct operation, or the numbers presented in the same sequence as the stem. She could not explain why she chose this as her answer, and it seemed she chose it randomly.

Cross-group comparison of the themes showed that most of the students in the *focusing on the sequence of numbers and ignoring the operation* theme were high or medium ability students (e.g., 34101_F, 35008_F, 94306_F, 94310_F, and 95224_F), while most students in the third theme were low ability students (e.g., 34245_M, 95230_F). It seemed that the high ability students might have been too adherent to their practice of looking for the numbers mentioned in the item in the exact same sequence and became too confident in selecting an answer without fully understanding the relationship mentioned in the item.

Item 20: Joba has some oranges. Dilu has 28 oranges which is 12 more than Joba. Which of the following number sentences is right for figuring out how many oranges Joba has?

- a. $28 + 12$
- b. $12 - 28$
- c. $12 + 28$
- d. $28 - 12$
- e. None of the above

Item 20 has subtraction operation with two two-digit numbers as its content. But this was made more complex because the stem mentions 'more' which usually triggers an addition operation. So, this requires some finesse from the students to determine that this should actually be a subtraction.

All the students were able to read the item without any trouble. From the interview responses, three prominent themes were identified. Similar to previously discussed items, the first theme was *phrase triggered responses*. Many of the respondents (nine in total) incorrectly answered this item because they decided this should be an addition based on the phrase 'more than' in the stem. They chose either "a" or "c" as their answer, both of which had the two numbers mentioned in the stem and an addition operation. While answering the item, students who were classified within this theme said things like "since the question said Dilu has 12 oranges more (saying more with emphasis) than Joba, it should be an addition. So, the correct answer is 'a', that is, $28 + 12$ " (35001_M) or, "'a', that is, $28+12$ is the correct answer because the stem mentions more" (35006_F). The second theme can be named as *beginning unknown*. Most of the students (ten in total) were able to look beyond the phrase 'more than' and understand that for this item, the beginning quantity was

unknown. So, even though it mentions ‘more than’ it should be a subtraction problem.

These students correctly chose “d” as their answer while saying things like, “d is the correct answer because we know how many oranges Dilu has and we know that is 12 more than Joba, so to get how many oranges Joba has, we need to subtract”

(25001_M) or, “d is the correct answer because Dilu has 12 oranges more than Joba, so to calculate the number of oranges Joba has, we would need to subtract 12 from 28 (as Dilu has 28 oranges)” (35008_F). The third theme was *confused random answer*.

One of the respondents (94310_F) got confused while answering this item. After considering the different alternatives for some time, she decided “e”, that is, none of the above was the answer. However, she could not explain why she thought this was the correct answer.

Cross-group comparison of the themes showed that most students who were able to go beyond the phrase triggered response and correctly answer the item were high ability students. However, there were some medium and low ability students who were within this theme as well.

Item 26: Rahat has 41 stamps, Rubin has 30 stamps. Lisa has 23 stamps more than what Rahat and Rubin have together. Which of the following number sentences is right for calculating how many stamps Lisa has?

- a. $41 + 30 - 23$
- b. $41 + 30 + 23$
- c. $30 - 23 + 41$
- d. $41 - 30 - 23$
- e. None of the above

All the interview respondents were able to answer item 26 correctly. The narratives from the interviews revolved around a single theme, *phrase triggered responses*. All the students seemed to be focusing on two phrases within the stem: more than and together. From these phrases, they determined that the appropriate number sentence should be an addition of the three numbers mentioned in the stem and they chose “b” as their answer. While answering the item they said things like, “b is the correct answer because Lisa had as many stamps as Rahat and Rubin together (emphasised) and then 23 more (emphasised)” (34102_F). or, “b is the correct answer because the question mentions *together* and *more*” (35001_M).

It was not possible to conduct any meaningful cross-group comparison for this item as students from all different ability groups had responded in the same fashion and fell under the same theme.

Item 32: Rubi got 2500 taka from her parents to buy new books. Nusrat got 700 taka less than Rubi and Rima got 300 taka less than Nusrat. Which of the following number sentences is correct for calculating the amount Rima got?

- a. $2500 + 700 + 300$
- b. $2500 - 700 + 300$
- c. $2500 + 700 - 300$
- d. $2500 - 700 - 300$
- e. None of the above

All the students were able to read Item 32 without any difficulty. Most of them (twelve) were able to answer it correctly in the interviews. There were three themes within the response narratives. The first theme was again, *phrase triggered responses* where students focused on the ‘less than’ phrase mentioned in the stem twice and decided that there should be two subtractions for the correct number sentence. They

looked at the alternatives and selected “d” as their answer. While answering, they said things like, “because Nusrat got 700 taka less and Rima got 300 taka less, it should be $2500 - 700 - 300$, so, d is the correct answer” (35001_M) or, “ $2500 - 700 - 300$ is the correct answer because Nusrat got 700 taka less and Rima got 300 taka less, so it should be subtractions” (25011_M). The second theme was *number focused responses* where some students (two) focused on the numbers mentioned in the stem and disregarded the appropriate operation. They considered addition to be the correct operation for this item and chose “a” as their answer. When asked why they thought this should be an addition, they could not explain. The third theme from the narratives was *confusion about appropriate operation*. Students whose response fell within this theme (five in total) felt there was some combination of addition and subtraction that was required for solving this problem, but they were not sure what the combination should be. As a result, they chose either “b” or “c” as their response. It seemed they were confused because there were three numbers involved. For example, one of the students (35008_F) looked at the alternatives for some time and said “c” was the correct answer. She read the stem again and suggested the number sentence should be adding the first two numbers and subtracting the last one. When asked why she thought this was the appropriate number sentence, she could not explain.

Cross-group comparison of the themes showed that most of the high and medium ability students’ responses were within the phrase triggered responses theme and they answered the item correctly. Only one student from the high ability group and one from the medium ability group were within the third theme.

6.2.1 Summary of analysis for Strand 2 items

Generally, the items in this strand worked well to assess students' ability to map conceptual relations to a numerical representation. Phrase triggered responses constituted a recurring theme for the items here. For all the items, students determine the correct answer when they were able to avoid the usual mistakes that happen while mapping conceptual relations to a numerical representation. For example, in some of the items (e.g., item 2, item 8), numerical representations consistent with the 'more than' or 'less than' phrases were the correct answers whereas there were some items (e.g., item 20) that were inconsistent (i.e., they had a phrase that usually triggers an addition operation but were actually subtraction). As a result, only students who were skilled in understanding the different representations were able to correctly answer both consistent and inconsistent problems.

The cross-group comparisons suggested that lower ability students were having difficulty to determine the correct answers for more difficult items because of their lack of skills in mapping the conceptual relation presented in the problem to a numerical representation. Overall, the response processes demonstrated in the interviews reflected the construct map hypothesised for this strand.

6.3 Strand 3 - Selecting an appropriate strategy for solution

There were six items in the test representing this area of AWPS. These items are: Item 3, Item 9, Item 15, Item 21, Item 27, and Item 33. Students' interview responses for these items will be discussed in this section.

Item 3: Solve this addition problem $2457 + 1319$. Show your work.

Item 3 consisted of adding two four-digit numbers with use of carrying once. Almost all the students answered this item correctly. All of them were able to read the item without any trouble and they used columnar addition (where they put one number under the other) to solve it. The major theme that emerged from the response data was students' use of *counting-up on finger sections* and *recalling addition results*. Most of them used counting-up on finger sections for the component additions. Moreover, there were some recalls used for some of the smaller component additions. Only one student (94306_F) made a mistake while carrying the number from the rightmost column and counted-up one more to get an answer of 3786. Cross-group comparison of the themes suggest that students from all ability groups had used counting-up and recall.

Item 9: Solve this subtraction problem $4869 - 1344$. Show your work.

Item 9 consisted of subtracting two four-digit numbers without borrowing. All the respondents in the interviews were able to read the problem without any trouble and all of them used columnar subtraction for doing the subtraction. There were two prominent themes in the narrative within the students' responses: *successful use of general subtraction strategies* and *failing to conduct the subtraction*. Most of the students could do the subtraction successfully as it did not require any borrowing. They used a mix of recall and counting-up or counting-down for the smaller subtractions. A few students (four) were unable to do the subtraction successfully. Some of them made a mistake while counting whereas others made a mistake by doing a borrowing where it was not applicable (25031_M). It seemed that the ones who made mistakes were having difficulty dealing with the large numbers. Cross-group comparison of the themes did not show any meaningful difference among the different ability groups' responses.

Item 15: Solve this addition problem: $52758 + 25936$. Show your work.

Item 15 consisted of adding two five-digit numbers with carrying two times. All of the respondents in the interviews were able to read the item and did this as a columnar addition where they put one number under the other before adding them. The prominent theme within students' responses was *successful use of carrying over* as almost all the students were able to carry-over where required to solve the addition. While doing the component additions they used a mix of counting-up and recall techniques. Two students could not answer this item correctly. One of them (94306_F) made a mistake while carrying over but the other (34107_M) made a

mistake in one of the component additions. Cross-group comparison of the theme did not show any difference among the different groups' performing this addition.

This item assesses students' use of addition strategies, but it is likely that sometimes students would incorrectly answer this item because of mistakes in counting, even after using the strategies correctly. It might be said that the cognitive load of adding two large numbers where carrying over is used twice makes counting errors more likely. However, it would be better if strategies could be assessed completely independent of counting in future iteration of the test.

Item 21: Solve this subtraction $6867 - 2593$. Show your work.

Item 21 involved the subtraction of two four-digit numbers with use of borrowing once. All the students were able to read the item and they put it down as a columnar subtraction for solving it. Two themes were extracted from the students' responses in the interviews: *successful use of borrowing* and *inability to use borrowing*. A large proportion of the students (13) were able to correctly answer this item. They were able to use borrowing strategies necessary for this item and reached the correct answer. While doing the component subtractions, they used counting-up, counting-down and recall techniques. The other students (7) were unable to use the borrowing strategy correctly and consequently got an incorrect answer or had to abandon answering the item. Some of them made mistakes in borrowing while others simply avoided borrowing by subtracting the smaller number (which was on the top row) from the bigger number (which was on the bottom row) for the component

subtractions. This did not make any sense, but it seemed they were not familiar with or confident enough to use borrowing and resorted to doing subtraction the only way they knew. Cross-group comparison of the themes showed that most of the students who failed to employ borrowing strategies correctly were from lower ability groups.

Item 27: Solve this addition $40695 + 17582 + 10346$. Show your work.

Item 27 was a relatively complex addition where there were three five-digit numbers and three instances of carrying were required to solve it. All the students did this as a columnar addition. Two themes were prominent within the students' response narrative for this item. The first one was *confident use of carrying-over strategies* and the second one was *difficulty with carrying-over strategies*. Most students (13) were able to answer this item correctly. They used carrying-over three times to solve this addition and did not make any mistake or show any discomfort with the procedure. They used a mixture of counting-up and recall techniques for the smaller additions that they had to do. The second theme was difficulty with carrying-over strategies. Six of the seven students who could not correctly solve this item during the interviews could not do so because they had trouble with carrying-over. Some of them simply could not do a carry-over while others made mistakes while performing carry-over. Such mistakes included carrying the wrong number (e.g., 34107_M carried 2 instead of 1) and forgetting to add the carried number (e.g., 95230_F).

Cross-group comparison of the themes showed that almost all the students who could not use carrying over strategies were from medium or low ability groups.

Another interesting phenomenon that emerged from the narrative of this item was compared to the lower ability group students, many of the students from the high ability group used 'recall' more for the component additions. It is interesting because they did not use recall as many times for the component additions of the previous items. One reason might be that when they need to do a number of smaller additions, they resort to doing recall to get them done quickly whereas while fewer smaller additions are necessary, they rely more on counting techniques to ensure they are doing it right. However, exploring this particular phenomenon is beyond the scope of this research and probably should be investigated more in depth in the future.

Item 33: Solve this subtraction $4912 - 2385 - 1456$. Show your work.

Item 33 was subtraction of three four-digit numbers. This is a complex subtraction which could be solved in a few different ways. It should be mentioned that students have such subtraction problems in their textbooks. Three different themes were extracted from students' responses to this item. The most prominent theme was *attempting columnar subtraction*. Most students (10) tried to solve this as other subtraction problems by writing it down as a columnar subtraction. However, none of them was successful in completing the subtraction and either gave up mid-way (e.g., 34101_F, 95203_F, and others) or made mistakes in borrowing and reached an incorrect answer (e.g., 34102_F, 34107_M, 35003_M, 25001_M, and others). The second theme was *unsuccessfully using adding strategy*. Some of the students (5) described how they were familiar with such subtractions where they would add two

numbers before subtracting. However, they could not employ this strategy successfully. All of them added the first two numbers and then subtracted the third number from the result which gave them an incorrect solution (e.g., 35001_M, 25003_M, and others). They knew there was a different strategy for solving this subtraction, but they were not skilled enough to employ that strategy. The third theme was *doing two separate subtraction*. Only two students were able to successfully solve this subtraction and they used this strategy. They first subtracted 2385 from 4912, and then subtracted 1456 from the result. This meant they had to use borrowing strategy three times, but they still managed to correctly answer the item. There were a few students (3) who did not attempt to solve the item at all. They were very nervous and uncomfortable and could not decide how to approach solving the item. After they were given enough time to try, the interviewer moved on to the next item.

Cross-group comparison of the theme showed that there were students from all ability groups in all the themes. So, it was not possible to determine if students from certain ability groups preferred certain strategies for solving this item.

6.3.1 Summary of analysis for Strand 3 items

Analysis of the responses suggested that, in general, items from this strand were assessing students' ability to use different strategies, as intended. Students who could use carry-over and borrowing strategies could answer the items correctly and students who were answering incorrectly were (mostly) making mistakes in carrying or borrowing.

Cross-group comparison could not be done for some of the items (e.g., item 9, item 15) but for the items where meaningful comparisons were possible, the themes showed that students who failed to employ appropriate strategies were mostly from the lower ability group.

6.4 Strand 4 - Performing the calculations on the numbers

There were six items in the test representing this area of AWPS. These items are: Item 4, Item 10, Item 16, Item 22, Item 28, and Item 34. Students' cognitive interview responses for these items will be discussed in this section. It should be mentioned that though some of these items had some carrying over or borrowing involved, these were intentionally kept at a minimal level to make sure the items focused more on students' ability to calculate.

Item 04: $522 + 87 = ?$

- a. 527
- b. 609
- c. 435
- d. 509
- e. None of the above

Item 4 in the test consisted of adding numbers (one three-digits and the other two-digits). Almost all the students (seventeen) were able to do the calculations and answered the item correctly. All of them used columnar addition to solve the item. There were two major themes extracted from the student narratives: *successfully using counting and recall* and *making mistakes while using counting and recalling*. Most of the students answered the item correctly and used a mix of counting-up and

recall for the component calculations. The students who could not answer the item correctly (34245_M, 94306_F, 95230_F) made mistakes while counting-up or recalling. Cross-comparison among the themes showed that the students who made mistakes were mostly from the low ability group.

Item 10: $328 - 75 = ?$

- a. 253
- b. 403
- c. 263
- d. 273
- e. None of the above

Item 10 consisted of subtraction of two numbers where one number had three-digits and the other had two-digits. All the students used columnar subtraction to solve this item and almost all of them did the subtraction correctly. Only a few of them (three) made mistakes while solving this item. Similar to the previous item, there were two major themes extracted from the student narratives: *successfully using counting and recall* and *making mistakes while using counting and recalling*. Those who answered the item correctly successfully used a mix of counting (only verbally counting or using finger sections) and recall. Counting was used noticeably more, compared to the previous item. Those who could not answer the item correctly were also using counting and recall but they made mistakes in at least one of the component subtractions. Cross-comparison among the themes did not show any noticeable difference among the different ability groups.

Item 16: Which of the following is the closest to the result for solving $94 + 33$?

- a. 110
- b. 130
- c. 120
- d. 140
- e. None of the above

Item 16 consisted of adding two two-digit numbers and then another level of calculation where students would need to find the closest number to the result they got from the addition. Most of the students solved this as a columnar addition. Three themes were noticeable within the students' responses: *successfully calculating the addition and the approximation, making mistakes in the addition, and making mistake in the approximation*. About half of the students (9) were able to answer the item correctly. They could successfully do the addition and were able to find the correct answer by calculating the closest number to the result from addition. Students used a combination of counting-up and recall for the smaller component additions. After getting 127 as the result from addition, they looked at the alternatives and did small calculations to find which one was the closest to 127.

Students who represented the second theme made mistakes while doing the addition. They started to do it as a columnar addition and were also using a mix of counting and recall to do the calculations. However, they miscalculated one or more of the smaller additions and as a result gave an incorrect answer.

The final theme was where students did the addition without any trouble but could not correctly find the number closest to the result they got. They used counting and recall for the component additions, got 127 as their result but miscalculated while determining the number closest to 127 from the alternatives and chose an incorrect

answer. It is noticeable how adding the proximation to the item added an additional layer of difficulty to the item.

Cross-comparison among the themes showed that most of the students who made mistakes in finding the number closest to their answer were students from the low ability group.

Item 22: Which of the following is the closest to the result for solving $87 - 53$?

- a. 37
- b. 44
- c. 27
- d. 30
- e. None of the above

Item 22 involved a subtraction (without borrowing) of two two-digit numbers with an added level of calculation where the students would need to find the closest number to the result they got from the subtraction. Almost all the students used columnar subtraction to solve this item. In addition, the students' narratives showed how they used both counting and recalling for the component subtractions. Three themes were noticeable within the student responses: *successfully calculating the subtraction and the approximation*, *difficulty with the subtraction* and *difficulty with the approximation*. About half of the students were able to successfully calculate both subtraction and approximation to correctly answer the item. These students represent the first theme. They used counting and recall for the subtraction as well as for the approximation (e.g., 35006_F and 35008_F). Students in the second theme were not able to do the subtraction successfully. Some of them made mistakes while doing the subtraction (e.g., 34245_M) whereas some (e.g., 94366_F, 95230_F) could not even

start doing it and gave up. A number of students (6) represented the third theme where they were able to do the subtraction but were not successful in the next level of calculation. They subtracted the two numbers, found 34 to be the result but were unable to choose the number closest to 34 from the alternatives as they made mistakes in calculating which is the closest.

Cross-comparison among the themes showed that most of the students who had difficulty with subtraction were from the low ability group. Similarly, most students who had difficulty with approximation were from medium or low ability groups.

Item 28: Which of the following is the closest to the result for solving for $2137 + 251$?

- a. 2350
- b. 2360
- c. 2420
- d. 2400
- e. None of the above

Item 28 consisted of adding (without carrying) two numbers, one with four-digits and the other with three-digits. In addition, approximation was required to determine the correct answer among the alternatives from the addition's outcome. Students' responses for this item showed that all of them solved it as a columnar addition. There were three themes within the response narratives: *successfully calculating the addition and the approximation*, *difficulty with addition* and *difficulty with approximation*. Most of the students (12) were in the first theme where they showed they could solve the addition and correctly chose the number closest to the addition's result from the alternatives. These students mostly used recall for the component

additions. The second theme represented students who had difficulty with the addition. Only a handful of students (3) were in this category. They made mistakes during the calculation or skipped the item because they did not know how to solve it (e.g., 95230_F). Students in the third theme were able to do the calculations necessary for solving the addition without any trouble but they incorrectly calculated the number closest to the addition's result. There were five students representing this theme.

Cross-comparison among the themes showed that most of the students in the third theme were from the medium or low ability groups. Students in the second theme were from all three different ability groups. It seems that the approximation calculation made it more difficult for the students in the lower end of the ability spectrum.

Item 34: Which of the following is the correct answer for solving $7322 - 3980$?

- a. 3342
- b. 4342
- c. 3242
- d. 4242
- e. None of the above

Item 34 was comprised of subtraction of two four-digit numbers with borrowing. No approximation was required for answering this item. The students' responses showed two major themes: *successfully calculating the subtractions* and *difficulty with calculating the subtractions*. About half of the students (11) were in the first theme. They were able to calculate the component subtractions within this subtraction using mostly counting and some recall. It was necessary to use borrowing in the calculation and they successfully used that. All the other students were in the second theme

where they had difficulty with calculating the component subtractions. Some of them made mistakes while doing the calculations, whereas a few others were too uncomfortable while solving this item and skipped it entirely (e.g., 95230_F, 94366_F).

Cross-comparison among the themes showed that most of the students who were in the second theme, that is, who could not use the necessary calculation to answer the item correctly, were from medium or low ability groups.

6.4.1 Summary of analysis for Strand 4 items

From the analysis it could be commented that the items in this strand are assessing students' ability to perform calculations. Students incorrectly answered items when they made mistakes in calculations and their responses did not show any other factors that contributed to their answering the items correctly or incorrectly. For the items that had an approximation as the answer, it was noticeable how there were two separate themes – one for students making mistakes while doing the first layer of additions/subtractions and another for students doing the first layer correctly but making mistakes while doing the approximation.

The cross-group comparison of the themes showed that students who made errors while doing the calculations were generally from the lower ability group.

6.5 Strand 5 - Verifying the solution and communicating the results

There were eight items in the test representing this area of AWPS. These items are:

Item 5, Item 11, Item 17, Item 23, Item 29, Item 35, Item 37, and Item 38. Students' interview responses for these items will be discussed in this section.

Item 5: Suppose you were asked to add 235 and 54, and you found the result was 289. How would you write the solution down?

a. 235

$$\begin{array}{r} 54+ \\ \hline 289 \end{array}$$

b. 289

$$\begin{array}{r} +54 \\ \hline 235 \end{array}$$

c. 54+

$$\begin{array}{r} 234 \\ \hline 289 \end{array}$$

d. 235

$$\begin{array}{r} +54 \\ \hline 289 \end{array}$$

e. None of the above

Item 5 assessed students' ability to identify the correct solution and representation of result for a two-number addition (maximum three digits). All the students were able to read the item without any trouble and almost all of them answered it correctly.

There were two prominent themes within the student response narrative for this item: *identifying errors in all the distractors* and *failing to identify some errors*.

Students whose responses fell in the first theme (18 in total) were all able to identify the errors in the various distractors. They explained why each option could not be the

correct answer (e.g., the numbers were misaligned, misplaced or wrong) and chose “d” as their answer. Only two students’ response fell in the second theme where they were unable to identify some of the errors in the distractors and chose an incorrect answer. Both of these students chose “b” as their answer and when asked why they thought this was correct said, “because it has the largest number on top” (35006_F). Cross-comparison among the themes did not show any meaningful result for the themes as most of the students were within the first theme. However, both students incorrectly answering the item were from the medium-ability group.

Item 11: Suppose you were asked to subtract 84 from 439 and you found the answer was 355. How would you write the solution down?

a. 84

$$\begin{array}{r} - 439 \\ \hline 355 \end{array}$$

b. 439

$$\begin{array}{r} - 84 \\ \hline 355 \end{array}$$

c. 439

$$\begin{array}{r} 84 - \\ \hline 355 \end{array}$$

d. 355

$$\begin{array}{r} + 84 \\ \hline 439 \end{array}$$

e. None of the above

Item 11 consisted of presenting the solution and result for a subtraction of two-numbers (maximum three digits). All the students were able to read the item and all of them answered it correctly. Students’ narratives revolved around a single theme for this item: *identifying errors in all the distractors*. All of them chose the correct answer

and pointed out the mistakes in the distractors. The errors were: misplaced numbers, misalignment of numbers, and wrong operation. Cross-comparison for the theme did not generate any meaningful result as students from all ability groups were able to answer the item correctly and were in the same theme.

Item 17: Suppose you were asked to add 26, 2533, and 325. You found the answer was 2884. How would you write the solution down?

a.
$$\begin{array}{r} 2433 \\ 26 \\ +325 \\ \hline 2884 \end{array}$$

b.
$$\begin{array}{r} 2533 \\ 325 \\ -26 \\ \hline 2884 \end{array}$$

c.
$$\begin{array}{r} 2533 \\ 325 \\ +26 \\ \hline 2884 \end{array}$$

d.
$$\begin{array}{r} 2533 \\ 325+ \\ 26 \\ \hline 2884 \end{array}$$

e. None of the above

Item 17 assessed students' ability to identify the representation of solution and result for an addition that had three numbers (maximum four digits). All the students were able to read the item and almost all of them answered it correctly. Three themes were discerned from the students' narratives for answering this item: *identifying errors in all the distractors, failing to identify some errors and considering all the representations as erroneous*. Most students' responses were within the first theme.

They went through the alternatives and pointed out what was wrong with each of them except for the answer. The distractors had errors like wrong numbers, wrong operation, and misaligned numbers which these students were able to identify. Only one student's (94366_F) response was categorised within the second theme, who was not confident about the convention of writing the solution for such addition problems and chose "d" as her answer. For the third theme there was also one student (25031_M) who considered all the representations of the solution as erroneous and selected "e", that is, none of the above as his answer. He could not explain why he thought this was the correct answer. Cross-comparison of the themes showed that both students who failed to correctly answer the item were from the low ability group.

Item 23: Suppose you were asked to subtract 129 from 2613 and you found the answer was 2484. How would you write the solution down?

a.
$$\begin{array}{r} 2631 \\ -129 \\ \hline 2484 \end{array}$$

b.
$$\begin{array}{r} -129 \\ 2613 \\ \hline 2884 \end{array}$$

c.
$$\begin{array}{r} 2613 \\ -192 \\ \hline 2484 \end{array}$$

d.
$$\begin{array}{r} 2613 \\ 129- \\ \hline 2484 \end{array}$$

e. None of the above

Item 23 was about presenting the solution and answer for a subtraction of two numbers (maximum four digits). All the students were able to read the item without any trouble. Three themes could be identified from their response narratives: *successfully identifying the errors in the alternatives, failure to identify error in numbers, and focusing on the numbers and ignoring alignment*. Almost half of the students' (9) responses were represented by the first theme. These students were able to identify the errors in all the alternatives and selected "e", that is, none of the above as their answer. They identified all the mistakes such as error in number (2631 instead of 2613 in a; 192 instead of 129 in c), error in answer (2884 instead of 2484 in b) and misaligned numbers (d). Several students (9) were unable to identify erroneous numbers in the alternatives and chose them as answers. Their responses form the second theme. They understood how subtractions are written but chose "a" (6 students) or "c" (3 students) as their answers as they could not identify the wrong numbers there. While checking the different alternatives they focused on the operation and alignment of numbers but failed to notice that option "a" had 2631 instead of 2613 and option "c" had 192 instead of 129.

The third theme within the response narratives was *focusing on the numbers and ignoring alignment*. Some of the students (2) focused on the numbers being correct ignoring the position of numbers and chose "d" as their answer. They identified the numbers and the result correctly but failed to follow the convention of writing down subtraction.

Cross-comparison of responses within the themes showed that almost all the students in the second theme, that is, students incorrectly answering this item

because they did not notice the wrong numbers, were from the medium and low ability groups. Similarly, students in the third theme were also from these two ability groups.

Item 29: A teacher is buying some new furniture for his school. He bought two tables which are 7 feet long and weighs 25 kilograms each. He also bought 5 desks, each of which are 5 feet long and weighs 20 kilograms. The teacher calculated the total weight of the furniture to be 150 kg and decided a method of transportation. Which of the following alternatives show the correct calculation and an appropriate transportation method?

- a. Weight $25+25+25+25+20+20+20= 150$ kg; a rickshaw for transportation
- b. Weight $25+25+25+25+20+20+20= 150$ kg; a truck for transportation
- c. Weight $25+25+20+20+20+20+20 = 150$ kg; a truck for transportation
- d. Weight $25+25+20+20+20+20+20 = 150$ kg; a bus for transportation
- e. None of the above

Item 29 assessed students' ability to check how an addition problem has been solved and presented. In addition, it assesses their ability to determine the correct answer based on its suitability in real life situations. The students' responses to this item generated four themes: *attention to the solution and real-life situation, focusing on real-life situation ignoring errors in the solution, looking for exact numbers from the item in the solution* and *confusion with the item*. Half of the students (10) were able to both determine the correct solution and understand the real-life situation to identify the correct answer "c". They were able to identify that the other alternatives had wrong calculations or transportation options that were impractical. For example, respondent 35001_M said he chose "c" as his answer because the other alternatives either had a transportation option that was not feasible, or the calculation was wrong. The second theme (*focusing on real-life situation ignoring errors in the solution*)

emerged as some students (5) seemed to be focusing on the real-life situation only. These students did not check which of the alternatives had correct calculation and chose “b” as their answer. The third theme (*looking for exact numbers from the item in the solution*) was represented by only two students. These students were looking for all the numbers mentioned in the stem of the item and when they could not find them, chose “e” (none of the above) as their answer. For example, respondent 35008_F chose “e” as her answer and when asked why she thought this was the correct answer she said, “because the question had 7 and none of the alternatives had it”. The fourth and final theme (*confusion with the item*) comprised student responses where they seemed confused with the item and chose “e” (none of the above) as their answer. They seemed to struggle with the amount of information provided in the item and when they could not confidently select any of the alternatives, they chose none of the above to get through the item.

Cross-comparison for the themes did not generate any meaningful results as students from different ability groups were dispersed across the different themes.

Item 35: The headteacher of a school is planning a victory day celebration in her school and wanted each student to have a small Bangladeshi flag. These flags come in packets of 200s. There are 4 packets of flags in the school. The school has 656 students in total but 126 of them will not be present during the celebration. The headteacher calculated that there will be 530 students in the celebration and the school would not need to buy new flags. Which of the following alternatives show the correct calculation and result for the number of students and number of packets of flags required?

- a. $656 + 126 = 530$; 4 packets of flag
- b. $656 - 126 = 530$; 4 packets of flag
- c. $656 - 126 = 530$; 3 packets of flag
- d. $656 - 126 = 530$; 2 packets of flag
- e. None of the above

Item 35 was comprised of subtraction of two three-digit numbers and understanding a real-life situation. Three themes were identified from the student responses for this item: *understanding of the solution and the real-life situation, focus on solution disregarding the real-life situation, and difficulty and discomfort with the item*. About one-fourth of the students interviewed (6) could answer this item correctly. They represent the first theme as they could determine the correct answer based on their understanding of the appropriate solution and the real-life situation mentioned in the item. For example, respondent 35001_M selected “c” as his answer and explained how it was the correct answer because the school will need 3 packets of flags and pointed out the mistakes in the distractors. Other respondents within this theme also said similar things. The second theme (*focus on the solution disregarding the real-life situation*) emerged from the responses of students who could verify the correct solution for the addition presented in the item but did not take the real-life situation into account. They selected “b” as their answer. For example, respondent 25001_M selected “b” as his answer and when asked why he thought it was the correct answered he said because it has the correct calculation and the school has 4 packets of flags. Most of the students’ responses (11) were categorised in the third theme (*difficulty and discomfort with the item*) for this item. Some of these students had difficulty understanding the item (e.g., 34102_F), others could read the item but seemed confused while choosing an answer. They chose either “a” or “e” as an answer but could not explain why they thought it to be the right answer (e.g., 34245_M, 94310_F). A few of them became very uncomfortable with the item and after trying to find the answer for some time skipped the item (e.g., 25031_M, 94366_F, 95224_F, and 95230_F).

Cross-comparison of the respondents among the groups showed that most of the students who skipped the item were from the low ability group. When considering the *understanding the solution and real-life situation* theme, there was a mix of high ability and medium ability students.

Item 37: Pia was asked to solve this problem: 'Rashed started for his school from home at 10.40 am and reached the class at 11.30 am. How long did it take for him to reach his class?'

She solved the problem and found the answer was 50. How shall she write down the result?

- a. 50 am
- b. 50 hours
- c. 50 minutes
- d. 0.50 hours
- e. None of the above

Item 37 was designed to assess students' ability to communicate the answer to a problem in the correct format (based on their real-life knowledge). All the students were able to read the item without any trouble. Two themes were identified from the students' responses: *understanding of proper answer format based on real-life knowledge* and *difficulty with the answer format*. Most of the students (16) were able to correctly answer this item and their responses fell under the first theme. They knew how to tell time and how the result for this item should be written down. A few students (4) had difficulty determining the answer format here. They seemed to have difficulty in communicating time and chose one of the other alternatives. For example, respondent 34245_M chose D (0.50 hours) as his answer and could not explain why he thought this was the correct answer.

Cross-comparison of respondents within the themes suggests that there was a mix of students from all three different ability groups within the first theme. However, for the second theme, all the students were from either medium ability or low ability group.

Item 38: Leon was asked to solve the following problem:
'Jenny bought 5 candies for 12.5 taka. She gave the shopkeeper a 20-taka note. How much shall she get in change?'
He solved the problem and found the answer was 7.5. How should he write down the answer?

- a. 7 taka
- b. 7 taka and 50 poisa
- c. 7.5 poisa
- d. 7.5
- e. None of the above

Item 38 was designed to assess students' ability to communicate result of a solution in the correct format. All the students were able to read the item without any difficulty. Two themes were identified in the students' responses for this item: *understanding of proper answer writing format* and *focusing on the numbers disregarding the writing format*. Most of the students (13) were able to correctly answer the item. There responses form the first theme (*understanding of proper answer writing format*). These students chose "b" as their answer and were able to explain how this is the proper format because it had the taka and poisa separately (which are the Bangladeshi equivalents of dollars and cents). It seemed, some students (6) were not familiar with the convention of writing monetary changes in Bangla and they focused on the resultant number too much. Their responses form the second theme (*focusing on the numbers disregarding the writing format*). These students chose "c" or "d" as

their answer which had the number mentioned in the stem but were not in the correct format.

Cross-comparison of respondents within the themes suggested that all the students who were in the second theme and incorrectly answered the item were from the medium or low ability groups. However, there were students from all ability groups in the first theme.

6.5.1 Summary of analysis for Strand 5 items

The items in this strand were designed to assess whether students could identify the correct presentation of solution and answer. Some of the items (Item 29 and 35) had specific elements to determine if students could take into account the real-life situations presented. Furthermore, the two final items (item 37 and 38) respectively assessed their abilities to present time- and money-related answers in correct format. Overall, the analysis of the cognitive interview responses showed that the items were mostly successful in assessing students' ability to verify the solution and communicate the results. Whenever the students were selecting a wrong answer, it was because they were unable to verify the solution presented correctly, did not account for the real-life situation or did not understand the correct answer format. However, one of the items (item 35) seemed to be confusing for a number of students. They were struggling with the amount of information presented in the item.

Cross-comparison of the different ability groups suggested that most of the students who skipped items or chose an incorrect answer did so because they could not identify the errors in solution or answer format, were from the medium or low ability group.

Chapter summary

This chapter discussed the analysis of the qualitative data collected through cognitive interviews in the final phase of the study. Students' response processes for the items in the different strands of the hypothesised AWPS framework were analysed. Generally, the response processes were found to be consistent with what the items were designed to assess. Some items had some minor issues that were identified for revision in future iteration of the test.

Chapter 7: Findings and discussions

7.1 Reflections on the study

7.1.1 What were the objectives of this study and were they fulfilled?

The aim of this research was to develop assessment materials that will enable teachers to identify specific difficulties students face while solving arithmetic word problems. The objectives were to:

- a) develop a theoretical framework for arithmetic word problem solving using an iterative process(as discussed by Wilson, 2005) by:
 - b) constructing tasks aligned with the framework,
 - c) trying them out in both small- and large-scale studies,
 - d) investigating the results using qualitative and quantitative methods,
- and
- e) evaluating the success of that process by investigating the validity and reliability of the results.

Based on these objectives, the research questions for this study were:

- i. What would be a theoretical framework for arithmetic word problem solving?
- ii. What items can be used to assess various stages/steps of this theoretical framework?
- iii. How can this framework be used to identify students' difficulty in solving arithmetic world problem?
- iv. How valid and reliable is this process of identifying students' difficulty?

This research endeavoured to develop a theoretical framework for arithmetic word problem solving and determine the validity and reliability of this framework through collection of empirical evidence. It was posited that arithmetic word problem solving (AWPS) has a number of component skills that are necessary for successfully solving such problems. A literature review was conducted in order to develop a hypothesised theoretical framework of AWPS. This review took a 'broad to narrow' approach to have a better understanding of the area. Subsequently, the first step was understanding what a problem is, then the review moved on to problem solving, discussed historical background of the concepts of problem solving, importance of domain specific problem solving and reviewed mathematical problem solving. Next, the specific area of interest of this study, that is, arithmetic word problem solving, was discussed and, based on different extant frameworks of problem solving, mathematical problem solving, arithmetic problem solving and with support of literature on other relevant areas (e.g., strategies in solving arithmetic word problems and error analysis in written problem solving), a theoretical framework for assessing arithmetic word problem solving was proposed. This framework hypothesises that there are five elements within arithmetic word problem solving. Students need to master these component skills to successfully solve arithmetic word problems. As outlined in the literature review chapter, the five strands are:

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution

- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

A test was constructed reflecting these strands in order to collect empirical evidence about the validity and reliability of this framework. It is suggested that the process of developing and determining the validity and reliability of the test will result in establishment of evidence for the validity and reliability of the framework. There were two iterations of the test. First, a pilot study was conducted with a small sample (n=62). Based on the analysis of the pilot study data, a revised test was constructed for final data collection (n = 808).

A range of evidence was collected to ascertain validity of the test. Table 7.1 outlines the different types of validity evidence collected.

Table 7.1
Validity Evidence Collected for this Research

Type of Evidence	Source
Evidence based on internal structure	Rasch analysis of the test (the Wright map, item fit, response fit and so on)
Evidence based on relations with other variables	Textbook AWPS items in the test Mathematics and Bangla test scores from school tests
Evidence based on test content	Test design and specification while designing the test Expert opinionnaire
Evidence based on response processes	Cognitive interview of respondents

In addition to validity, reliability was calculated through WLE person separation reliability, EAP/PV reliability, and Cronbach's Alpha. The following sections briefly summarise the validity- and reliability-related findings from the final data collection.

7.1.1.1 Validity evidence based on internal structure of the test

This research adopted an Item Response Theory (IRT) approach to evaluate the internal structure of the test. A dichotomous Rasch analysis was used. The Wright map from item analysis demonstrated some minor divergence from the hypothesised difficulty of the items for the different strands. However, the analysis showed that the item difficulties were well distributed (from around -2.5 logits to 2.5 logits), item fit statistics suggested that almost all the items were within a reasonable range (weighted MNSQ 0.88 to 1.14) and did not show any extreme over-fit or under-fit. The fit t statistics for the items also indicated a good fit. The only exception was item 35, showing poor result in weighted MNSQ (infit) value (1.25), which was outside of the 95% Confidence Interval (0.9 to 1.1). Hence this item was recommended for removal in future iterations of the test. Respondent fit was also reasonable: only 3.5% students demonstrating underfit above 1.96 which is close to 5% – the expected value if all the respondents were fitting Rasch model and the data were randomly generated. All the items had a positive item-rest correlation value with most of them (27 out of 32) being above 0.20. Item 35 had the lowest item-rest correlation value (0.00) which strengthened the argument for its removal. Overall, the data showed good adherence to the Rasch model and can be considered to demonstrate reasonably good evidence of validity in terms of internal structure of the test.

The idea of using five separate strands for assessing arithmetic word problem solving was supported in a model comparison analysis. The analysis showed that a multidimensional model was a better fit for the data, compared to a unidimensional model. This suggests that the AWPS construct is multidimensional.

7.1.1.2 Validity evidence based on relations with other variables

Two types of data were collected with the aim of establishing validity arguments based on the AWPS test's relation with other variables. These were (a) scores from six textbook items included in the test that was administered and (b) students' test scores for Bangla and Mathematics from their schools. The expectation was that the students' performance in the AWPS test should be positively correlated with their performance in the textbook items and their performance in the school Bangla and Mathematics tests. The textbook items were considered as an existing assessment that determines their ability to solve arithmetic word problems. The Mathematics and Bangla test scores were considered as proxies of their Mathematics and Language abilities.

While determining the correlation between students' performance in the AWPS test and in the other assessments (i.e., score in textbook items, Mathematics and Bangla test scores), students' ability estimates for the AWPS test obtained in logit values from the Rasch analysis were converted to scaled scores using a linear transformation (where Scaled Score = Ability estimate in logit $\times$ 100 + 500). However, the data for the textbook items, Mathematics test and Bangla test were simply the raw scores.

Comparison of the students' total scores in the textbook items and their scaled score in the AWPS test via Pearson correlation coefficients demonstrated a positive correlation (0.62) between the two scores and this was statistically significant at the 0.01 level (two-tailed). Furthermore, there was a statistically significant positive correlation between these two scores when grade 4 and grade 5 students were analysed separately.

Comparison of the students' scaled scores in the AWPS items and their mathematics scores in school examinations also yielded a positive Pearson correlation coefficient (0.40). Similarly, scaled scores in the AWPS items and Bangla scores in school examinations were also positively correlated (Pearson correlation coefficient 0.49). Both these were statistically significant at 0.01 level (two-tailed). It is interesting to note that the correlation coefficient was higher for the comparison of the AWPS test with the Bangla test scores. However, it was not possible to delve deeper into this as the test items for the school administered tests were not available.

Students' performance in the sub-domains of the AWPS test was also compared with the textbook item scores and Mathematics and Bangla test scores. A multidimensional Rasch model was utilised (in *ConQuest*) to calculate students' ability estimates in the different dimensions (representing the sub-domains). The respondents' ability estimates in the different sub-domains were found to be positively correlated with the textbook item scores and Mathematics and Bangla test scores. The correlations were found to be statistically significant at the 0.01 level (2-tailed). This suggests, there is evidence of validity for using the separate sub-domains of the AWPS construct. However, the correlation coefficient values for the sub-

domains were small compared to the coefficients calculated for the whole AWPS test. This is likely due to the small number of items in each sub-domain.

Overall, the AWPS test showed good evidence of validity in relation to other variables.

7.1.1.3 Validity evidence based on test content

There are two types of test content-based validity evidence that were accumulated for this research. One was the use of a rigorous process that was followed to develop the test specification and the items. The evidence for this process is less formal and was not quantified. The other source of evidence was experts' views about the test items collected through an expert opinionnaire. The data from this opinionnaire was analysed using descriptive statistics, and it is discussed below.

Test design and specification

Following predefined and research supported procedures while designing a test undergirds the validity argument for the test. This study followed the four building blocks outlined by Wilson (2005) as steps of an iterative process. The first two building blocks, that is, identifying and defining the construct and developing items, are relevant to validity evidence based on test content. Observing recommendations of Wilson (2005), Wolfe and Smith (2007), Wu et al. (2016) and other researchers and after reviewing the Bangladeshi primary curriculum, textbooks and different mathematics tests, this study developed construct maps and item specifications for the pilot test (Appendix 3.1), which was later revised for the final test (Appendix 3.2).

Expert opinion about the test items

The constructs and test specifications mentioned earlier were reviewed by experts and the items were written based on those specifications. In addition, the item development activities included writing several draft sets of items and paneling them with mathematics education experts from the University of Melbourne. Their expertise was also useful in revising the pilot test, so a more refined test could be developed for the final data collection. However, after the final set of items was constructed, formal data were collected from Bangladeshi experts about the suitability of the items for the AWPS framework. Analysis of this data showed that overall, the experts' association of strands with items agreed with what was hypothesised, that is, they nominated items for strands they were hypothesised to be representing in this research. All the items obtained the highest percentage for the strands they were representing. However, there were some items where a relatively large portion of the experts chose a strand different from what it was for. Another interesting finding was how the responses were sharply in line with the item design for the items in the early part of the framework (i.e., items chosen for the strand they were hypothesised to be representing) and gradually became more dispersed for the later strands. This could be evidence that arithmetic word problem solving is a process and consequently, for items in the later strands (e.g., verifying the solution and communicating the results), the respondents felt that the previous strands were coming into play to some extent (e.g., comprehension of the problem).

7.1.1.4 Validity evidence based on response processes

A data collection method that collects insights about how the respondents responded to the items is an important part of item development. Such evidence can be used as a summative tool for gathering validity evidence along with their formative input in the item design process (Wilson, 2005). This research conducted cognitive interviews with 20 respondents from three schools to collect validity evidence based on the response processes. The idea was to check whether the items were assessing what they were designed to assess. One aspect of determining this was to check whether lower ability students were struggling with the more difficult items and what seemed to be causing difficulties for them. Data from these cognitive interviews were analysed using a qualitative method called 'reduction and synthesis' (Miller et al., 2014).

Analysis of the qualitative data showed that the items were mostly working as intended. For the 'Comprehension of the Problem' items, it was noticeable how comprehension is different from reading. Most students were able to read the items (stems and distractors both) without any trouble, yet they could not comprehend the meaning in order to select the correct answer. More complex items (items that were designed to be more difficult) seemed to be harder to answer for some students, demonstrating the difference of abilities required for correctly answering those items.

For the items in the 'Mapping conceptual relations to a numerical representation' strand, it was noticeable how the students' ability to correctly map the problem into a number sentence was influenced by their ability to correctly interpret phrases like 'more than' and 'less than' in the stem. Most students were able to correctly map the problem when the phrases were consistent with the operation

(i.e., 'more than' meant an addition operation) but lower ability students struggled when the phrases were inconsistent with the operation (i.e., the stem contained 'more than' but the required operation was subtraction). Previous research (e.g., Fuson et al., 1996; Lewis & Mayer, 1987) has suggested that there is a progression of schema use where at the initial levels the solvers depend on the schemas triggered by the phrases in the problem sentence and gradually get skilled in transforming the inconsistent problems before understanding how to solve them. This seemed to be the case for respondents of this research as well.

For the third strand, that is, 'Selecting an appropriate strategy for solution' the items included in the test focused exclusively on whether students could select strategies necessary to do addition and subtraction. The items mainly tapped into their ability to use carry and borrow strategies for addition and subtraction. Analysis of students' qualitative data showed that students who were able to use these strategies sufficiently were able to answer the items and those who could not employ these strategies could not answer the items. It was noticeable how the more difficult items that required use of carry/borrow more than once had some students struggling and making mistakes. However, for some students this cognitive load led to mistakes in calculations and, consequently, an incorrect answer. The interviews also showed how the students used calculation techniques like recall, counting-up, counting-back and so on. It was interesting to note that compared to the lower ability students, students from the high ability group used 'recall' more for the component additions of the larger additions in the later items. They did not use recall as many times when the addition was small for the initial items. One reason for this might be because when they do a number of smaller additions (as components of a larger addition), they

resort to doing recall to get them done quickly whereas when fewer smaller additions are necessary, they rely more on counting techniques to ensure they do not make a mistake. Or they might just be doing that out of habit. However, exploring this particular phenomenon was beyond the remit of this research and could be investigated more in-depth in the future.

Items in the fourth strand, that is, 'Performing the calculations on the stored numbers' were pretty straight forward. Students could either do the addition/subtraction or they could not do it. However, it was noticed in the analysis that some students struggled with the items where they were asked to do an approximation after calculating the answers. This was an added level of calculation that they had to do in their minds, and some had difficulty doing that. In addition, it was noticeable that the more complex items (that had addition/subtraction of larger numbers) troubled the lower ability students more.

Most of the fifth strand (verifying the solution and communicating the results) items consistently showed that students who could identify the errors in presentation of results, could answer the items correctly. There were many types of errors presented in the distractors and only students who understood the different types of errors (e.g., misplaced numbers, misalignment of numbers, wrong symbol for operation, etc.) could successfully answer the items. The items that focused on assessing students' real-life understanding of a situation also seemed to work well, as those who did not consider the real-life scenario presented in the item ended up selecting an incorrect answer. However, some items seemed to be dependent upon

students' ability to read and comprehend the information (e.g., Item 35). That should be considered in future iterations of the test.

A recurring theme from the analysis of the items from different strands seemed to be that, generally, higher ability students (who had a better merit position in their schools) seemed more comfortable in answering the more difficult items and lower ability students seemed to be struggling to answer them.

7.1.1.5 Reliability

Test reliability was calculated using both IRT and CTT measures. The WLE person separation reliability was 0.81 and the EAP/PV reliability was 0.82. The Cronbach's Alpha value was 0.82.

All the different measures of reliability demonstrated a coefficient value of above 0.8. Generally, a reliability value of over 0.8 can be regarded as indication of a highly reliable assessment.

In addition, reliability of the different sub-domains of the AWPS test was calculated using *ConQuest*. The EAP/PV reliabilities for the different dimensions (representing the sub-domains) were high (≥ 0.70). However, the WLE person separation reliabilities for the different dimensions ranged between 0.28 to 0.47, which suggest that the items in the sub-domains had low to moderate sensitivity in distinguishing students of different ability. If the sub-domains of the AWPS test are administered as separate tests in future, one would need to increase the number of items for each to improve the WLE person separation reliability.

7.1.2 Significance of the study

The significance of this study can be discussed from three perspectives. First, from a global perspective, this research adds to the growing body of research that highlights the componential nature of mathematics. Future research can replicate this study in different settings to refine the AWPS framework and use it in different contexts. The methodology used in this study, and the exemplary materials that have been produced can be used in future research to develop (a) teacher professional development procedures and materials that can be used to help teachers adopt this new approach, and (b) assessment materials and teacher professional development materials for other areas of mathematics, including mathematical problem solving (e.g., arithmetic word problem solving where other operations like multiplication or division are used).

Second, in the Bangladeshi context, this research adds to the limited amount of research that has been done in primary mathematics. Problem solving in general and mathematical problem solving in particular is a vastly under-researched area in Bangladesh. Based on this study, future researchers might find it to be an interesting area for further research. In addition, the findings of this study could inform classroom practices and lesson designs in the country.

Third, this research is significant for the researcher (i.e., me). This enabled me to learn new methods and methodology of research and assessment. I was able to use my previous experience as a teacher and a researcher in a new context and learned many new things along the journey. This research facilitated the expansion of my

knowledge horizon on assessment techniques and mathematics education. Hopefully, this experience will pave the way for further development of knowledge and skills in the future for me.

7.1.3 Successes and limitations of the study

One of the successes of this study was to collect data from a large sample of students in Bangladesh despite initial setbacks (e.g., school closures). The test was administered with 826 students in nine different schools (808 students were retained for the final analysis). The schools were quite diverse in nature, that is, it was possible to include different types of schools (e.g., government primary schools, privately managed primary schools, high school attached primary schools) in the study. In addition to the test administered in the schools, students' Bangla (language) and Mathematics test scores in school examinations were collected. During the data collection there were widespread school closures due to heavy monsoon rain and flooding in some areas. Even with such drawbacks, the schools participated in the study with the data collection time extended. In retrospect, this was possible for two reasons: (a) communicating with the schools long before the actual data collection started and (b) developing a personal relationship with some schools at the initial stage of the data collection, which helped in getting other schools onboard. These came as a learning point for the researcher in that maintaining good communication and developing personal relations can help in data collection when adverse situations arise. In addition to administering the test and collecting students' school examination test scores, 20 think-aloud interviews were conducted with students and expert

opinion data were collected from 24 experts. Most of the expert opinion data were collected through email communications.

Another success of the study was the retention rate for responses in the research. Initially, data were collected from 826 students in the final phase of the study. After cleaning the data for incomplete test scripts and other anomalies within the responses, 808 respondents were retained for the final analysis. That is, almost 98% of the respondents were retained for the analysis. This retention rate is very good. One major reason for this happening was probably the fact that the expectations from the students were clearly explained before the test and they knew they were taking the test as part of a research study and this had no stake in their academic record within the school. Explaining the research and the test to the schoolteachers and having them assist with the test administration also helped in putting the students at ease.

The retention rate of items can also be considered another success in this study. During the piloting stage, there were 34 items based on the AWPS framework. After analysing the pilot data, 32 of these items were retained and used (following some revision) in the final test. After analysing item data for the final test, only one item (item 35) was considered to be problematic and it was suggested that this item be revised or deleted in future iterations of the test. So, from the initial 34 items, 31 (91%) have been retained. That is quite remarkable. It is likely that the rigorous planning stage for writing the items and revising multiple drafts before finalising items for the pilot study helped in writing some quality items that had a higher possibility of retention.

The use of non-probability sampling technique is a limitation of this study. For selecting schools, a convenience sampling was used and then whole classes of grade 4 and grade 5 students were selected from the schools. Although the schools were quite diverse (government, non-government, etc.) and had students from different socio-economic backgrounds, they were all from a single district in Bangladesh. Therefore, the student respondents cannot be considered as a representative sample of Bangladeshi students. This limitation should be kept in mind while considering the findings of this study.

7.1.4 How could this work be extended, and what could be done differently in future studies?

The test designed for this study was broad, as it had to accommodate the five strands of AWPS and be suitable for the two different grades of students taking the test. In addition, it included both addition and subtraction operations. Future iteration of this research could focus on a single grade and a single operation (either addition or subtraction) which would make designing and administering the test easier. Use of the test in schools or classrooms could also focus on a single grade or a single operation and could also consider assessing some of the sub-domains/strands instead of all.

Future studies on teachers' professional development could utilise the findings of this study to explore teachers' understanding of the different areas of AWPS. This in turn could contribute to developing professional development materials for mathematics teachers based on this work.

Collection of validity evidence through the expert opinionnaire is another area that could be done differently in future research of a similar kind. The opinionnaire developed and administered in this research had a single response for each of the items. That is, the experts were given an item and asked to select one strand from the five options that they thought the item was most suitable for. In future research, this could be done slightly differently where the experts are asked to select more than one strand for an item. They could be asked to rank the strands they thought an item was most suitable for. In such a design, an item would be presented, and the respondents would be asked to rank among the five strands based on their opinion about that item's suitability for each strand. This would give them an opportunity to demonstrate whether they thought an item represented more than one strand (which is likely). Another change regarding the expert opinionnaire could be the number of respondents. In this research, data were collected from 24 experts. Future research with better resources could target getting a larger sample of diverse respondents for this validity evidence. In addition to allowing more robust analyses, this could provide an opportunity of comparing opinions from different groups of experts.

Another thing that could be done differently in future iteration of this research is related to the strategy items. There are a number of strategies for addition and subtraction that students use consciously or unconsciously while doing addition and subtraction. Such strategies include counting up, counting back, doubling, rounding up, and so on. In some countries (e.g., in Australia), there are resources that explicitly help students to learn these strategies. But that is not the case in Bangladesh. As a result, students were not familiar with these strategies and could not answer the items in the pilot test that represented these strategies. Due to students' confusion

and discomfort with the items that asked them about this type of strategy in the pilot, the items were removed in the final test. However, in future research, based on suitability of the context, these could be reintroduced. For example, if this research is conducted in Australian schools in future, it could be possible to include the strategy items used in the piloting to check how feasible it is to assess students' use of such strategies.

7.1.5 What changes could be brought to the methodology of this study?

This research used an item response theory (IRT) approach for the psychometric analysis of the test data. Rasch's dichotomous model, a one parameter IRT model (item difficulty parameter only), was considered appropriate for the analysis. In the future, classical test theory (CTT) approach could be considered for additional analysis. "There are theoretical differences between IRT and CTT" (Wu et al., 2016, p. 26) in making sense of the data collected about the measured construct. In CTT, a person's true score (which can theoretically be measured by multiple parallel tests) shows the level of construct s/he has achieved whereas, in IRT, an explicit latent trait (the construct) is defined through the model and an individual's scores and item responses reflect her/his position or level in the construct (Wu et al., 2016).

IRT is generally considered superior and has certain advantages over CTT, for which it was the preferred approach in this research. Some advantages of IRT over CTT are as follows: IRT has a more item-level (micro) focus compared with CTT, which has a test-level (macro) focus, validity is unaffected in IRT if some items are omitted from a test, while in CTT validity is for a whole test and deletion of any items invalidates the

whole test; IRT generates interval scale data whereas CTT generates ordinal data; ability estimate is sample-independent in IRT but observed score of CTT is sample dependent, and so on (Kean & Reilly, 2014a, 2014b). Overall, IRT is more robust than CTT. However, an advantage of CTT over IRT is the familiarity of the CTT to a wider audience as IRT is newer and some researchers/readers (especially in Bangladesh) may find it unfamiliar. As such CTT-generated measures can be used as a complement to the IRT measures for future research or publications.

Although IRT was used for determining the psychometric properties of the test, some raw score-based analysis was also conducted in this study. For example, in addition to calculating WLE person separation reliability and EAP/PV reliability (which are part of the IRT approach), Cronbach's Alpha value was calculated for determining the test's reliability (which is based on raw scores).

For the qualitative analysis of the cognitive interviews, although the ability levels from the construct map were implicitly addressed during the analysis, the focus was more on the items. Future analysis of cognitive interviews could be done differently where the ability levels are more explicitly the focus.

7.1.6 What other types of analysis could be conducted?

It was not possible to conduct Differential Item Functioning (DIF) analysis for the items in this research. DIF analysis determines whether different groups of students are responding differently to an item (Wu et al., 2016). This helps in determining if there is performance difference among groups based on any characteristics other than their level of the trait being measured. The primary reason for choosing not to run DIF analysis was the fact that the sample of students was not diverse enough and there

were not enough resources to undertake it. However, a follow-up paper could be written from this study where boy/girl DIF analysis is conducted and that could be used to illustrate the techniques of DIF for educational researchers. In addition, future researchers could collect data from a more diverse sample and collect more background data about the students which would enable them to run DIF analysis based on other contextual factors.

7.2 Implications of this research for Bangladesh

This section discusses how the findings from this research can be used by different stakeholders of primary education in Bangladesh. These stakeholders include the policymakers and mathematics teachers.

7.2.1 Policymakers

It has been widely accepted in Bangladesh that the current primary education system has shown huge improvements in terms of access to education (e.g., net and gross enrolment) and other quantitative parameters (e.g., gender equity) but there remain some major issues around the quality of education. The most recent sector-wide intervention for primary education in Bangladesh, the fourth Primary Education Development Program (PEDP 4), identifies quality as one of the key focal points and recommends review and revision of the existing curricula, along with textbooks and teaching-learning materials (World Bank, 2018a). As such, PEDP 4 allocates considerable resources for revision of the curriculum as part of improving the quality

of primary education. This process of revising the primary curriculum is in operation now and the plan is to gradually roll out the new curriculum and textbooks in the coming years (Akhter, 2018; Amin, 2020).

While reviewing the current mathematics curriculum this study identified that there was no indication of how assessment is to be conducted in the curriculum, that is, the curriculum does not mention anything about what assessment should look like for the subject. This could be taken into account while revising the curriculum by including some discussion on assessment.

The empirical findings from this study could be used to inform the revision of the curriculum as well. As has been identified previously by different researchers, although the primary curriculum in Bangladesh is competency-based, there is a scarcity of competency-based assessments. Most of the assessments in practice revolve around rote learning and recall of textbook contents. The findings from this study could be helpful in designing assessment that focuses on students' skills of mathematics, as well as the content. As demonstrated by previous competency-based assessments in Bangladesh (e.g., the NSA, the Education Watch reports), and to some extent by this research, students do not perform well when they are taking a test that taps into the skills they are supposed to demonstrate instead of a test where the items are from their textbooks. The framework that has been validated through this study could be used to focus on the different skills of arithmetic word problem solving and work as a suggestive framework for writing about assessment in the curriculum and the teachers' guide/teachers' edition.

Assessment is outlined in the teachers' guides (which are called teacher editions in Bangladesh) for primary mathematics. A review of the assessments outlined for arithmetic word problem solving (for addition and subtraction operation) suggest that the focus of assessment in these teacher editions was on determining whether the answer to the word problem was correct/incorrect. There was no advice regarding how to determine if the students were having trouble with any specific aspect of solving the word problem. Though the current curriculum highlights the necessity of having provisions for ongoing assessment of learners' progress and provisions for remedial teaching to address the deficiencies of learning, there is little or no guidance on how to do this in the teacher guides. Based on the findings of this study, future revisions of the teacher editions could include an outline of the different elements that come into play while solving word problems and how it is necessary to understand which element students are having trouble with instead of focusing only on correct/incorrect answers, and how to design assessment that determines with which element students are having trouble. This could help teachers to plan remedial teaching for the areas where their students are found to be having difficulty.

As it was previously mentioned, the teaching support material (teachers' guide/ teacher edition) for the current curriculum was not available for the teachers till 2017 even though the curriculum has been in operation since 2012. Furthermore, the textbooks and other teaching learning materials associated with the curriculum have never been systematically evaluated (World Bank, 2018a). Therefore, another recommendation would be to develop the teacher support materials concurrently while revising the curriculum and textbooks and make these available for the teachers while the revised curriculum is gradually rolled out.

7.2.2 Mathematics teachers

This study has shown the validity and reliability of the proposed framework for Arithmetic Word Problem Solving (AWPS). This framework could be used by teachers to assess their students' arithmetic word problem solving ability in the classroom. The items designed for this study focused on grade 3 and grade 4 contents to have a good spread of content in the test. However, teachers could use the framework to develop their own items focusing on one grade. In addition, they could design items for addition or for subtraction only. The framework could be used for diagnostic purposes too, where they do not solely focus on the correct/incorrect answers but look into the areas where students might be struggling. For example, a teacher could develop a set of items based on all five strands of the framework to take a diagnostic test of students starting grade 4. That would give the teacher an idea of what the students' strengths and weaknesses are. At later stages of teaching, the teacher could then develop items on each individual strand and check whether classroom teaching has successfully alleviated students' weaknesses identified in the initial test.

The AWPS framework developed in this study could be used as a progression for Arithmetic Word Problem Solving (addition and subtraction problems) and could be used by teachers in conjunction with other methods of conducting assessment and giving feedback/designing follow up instructions. For example, a teacher could develop a test based on this framework and put the students' performance on a Guttman chart (as discussed by Griffin, Francis, & Robertson, 2018). They could then

identify the students' Zone of Actual Development (ZAD) and Zone of Proximal Development (ZPD) and identify ZPD groups for future instructions.

In general, this framework could give teachers an understanding of how some knowledge and skills are consolidation/summation of component skills and knowledge. Working with this framework, they might acknowledge how solving word problems requires students to use a range of different skills (the five strands) and how having trouble with one of those skills might mean they are unable to solve the problem. This in turn could help them think about other mathematical concepts and look for the component or prerequisite skills for those concepts. This means they will then be able to think about how to assess the consolidated skill along with the component skills so the students can be better taught, and students' difficulties could be better addressed in future teaching.

The materials developed in this study could be used as a basis for developing professional development materials for teachers. Furthermore, future research based on the findings of this study could help in developing assessments and professional development materials for other areas of mathematics.

Chapter summary

This is the final chapter of the thesis. It has recapped the findings of the study reflecting on the objectives and how they were fulfilled. In addition, the significance of the study, its successes and limitations were outlined. Furthermore, this chapter illustrated how this work could be extended and what could be done differently in future studies, what changes could be brought to the methodology, and what

additional types of analysis could be conducted. Finally, the implications of this research for the policymakers and teachers in Bangladesh were discussed.

Bibliography

- Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, 31, 162-172. doi:10.1016/j.stueduc.2005.05.008
- Adams, T. L. (2003). Reading mathematics: More than words can say. *The Reading Teacher*, 56(8).
- Ahsan, S. (2009). Classroom assessment culture in secondary schools of Dhaka city. *Teacher's World*, 34(1), 231 - 244.
- Akhter, S. (2018, December 1). Govt to revise school curriculums after 7 years. *NEWAGE*. Retrieved from <http://www.newagebd.net/print/article/57573>
- American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.
- American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2014). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.
- Amin, M. A. (2020, February 4). Education curriculum, textbooks to go through massive revamp. *Dhaka Tribune*. Retrieved from <https://www.dhakatribune.com/bangladesh/education/2020/02/04/education-curriculum-textbooks-to-go-through-massive-revamp>
- Anderson, J. R. (1990). *Cognitive psychology and its implications* (3rd ed.). New York: Freeman.
- Anderson, J. R. (2015). *Cognitive psychology and its implications* (8th ed.). New York: Worth Publishers.
- Artz, A. F., & Armour-Thomas, E. (1992). Development of a cognitive-metacognitive framework for protocol analysis of mathematical problem solving in small groups. *Cognition and Instruction*, 9(2), 137-175.
- Asadullah, M. N., & Chaudhury, N. (2013). *Primary schooling, student learning, and school quality in rural Bangladesh (working paper 349)*. Retrieved from <https://www.files.ethz.ch/isn/175734/primary-schooling-student-learning-school-quality-rural-bangladesh.pdf>
- Ashcraft, M. H. (1990). Strategic processing in children's mental arithmetic: A review and proposal. In D. F. Bjorklund (Ed.), *Children's strategies: Contemporary views of cognitive development* (pp. 185-211). Hillsdale, NJ: Lawrence Erlbaum.
- Asian Development Bank. (2011). *Bangladesh: Second primary education development program*. Retrieved from <https://www.adb.org/publications/bangladesh-second-primary-education-development-program>
- Australian Curriculum Assessment and Reporting Authority [ACARA]. (2016). *The Australian curriculum: Mathematics*. Retrieved from <http://www.australiancurriculum.edu.au/download/>
- Australian Education Council. (1990). *A national statement on mathematics for Australian schools*. Retrieved from <https://files.eric.ed.gov/fulltext/ED428947.pdf>
- Australian Education Council. (1994). *Mathematics - a curriculum profile for Australian schools*. Retrieved from <https://research.acer.edu.au/cgi/viewcontent.cgi?article=1000&context=mathematics>
- Azim, F. (2018). An analysis of the secondary school certificate examination: The case of creative questions. In R. Chowdhury, M. Sarkar, F. Mojumder, & M. Roshid (Eds.), *Engaging in educational research. Education in the Asia-Pacific region: Issues, concerns and prospects* (pp. 221 - 238). Singapore: Springer.

- Bahar, A. (2013). *The influence of cognitive abilities on mathematical problem solving performance*. (PhD Thesis), University of Arizona, Arizona.
- BANBEIS & MWCA. (2005). *Report on primary education in Bangladesh: Challenges and successes*. Dhaka: The Third Summit on South Asian Development: South Asian Association for Regional Cooperation (SAARC).
- Banu, L. F. A. (2009). Problems and misconceptions facing the primary language education in Bangladesh: An analysis of curricular and pedagogic practices. *BRAC University Journal*, 1(1), 1 - 10
- Bassok, M., Chase, V. M., & Martin, S. A. (1998). Adding apples and oranges: Alignment of semantic and formal knowledge. *Cognitive Psychology*, 35(2), 99-134.
- Beatty, P. (2004). The dynamics of cognitive interviewing. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods of testing and evaluating survey questionnaires* (pp. 45-66). Hoboken, NJ: John Wiley & Sons.
- Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71(2), 287-311.
- Bergqvist, E., & Österholm, M. (2010). *A theoretical model of the connection between the process of reading and the process of solving mathematical tasks*. Paper presented at the MADIF 7, The Seventh Mathematics Education Research Seminar, Stockholm.
- Best, J. W., & Kahn, J. V. (2014). *Research in education* (10th ed.). Essex, England: Pearson.
- Binder, T., Sandmann, A., Sures, B., Friege, G., Theyssen, H., & Schmiemann, P. (2019). Assessing prior knowledge types as predictors of academic achievement in the introductory phase of biology and physics study programmes using logistic regression. *International Journal of STEM Education*, 6(33). doi:<https://doi.org/10.1186/s40594-019-0189-9>
- Bisanz, J., & LeFevre, J.-A. (1990). Strategic and nonstrategic processing in the development of mathematical cognition. In D. F. Bjorklund (Ed.), *Children's strategies: Contemporary views of cognitive development* (pp. 213-244). New Jersey: Lawrence Erlbaum Associates.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7-74.
- Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 92(1), 81-90.
doi:<http://dx.doi.org/doi:10.1177/003172171009200119>
- Blum, W., & Niss, M. (1991). Applied mathematical problem solving, modelling, applications, and links to other subjects - State, trends and issues in mathematics instruction. *Educational Studies in Mathematics*, 22(1), 37-68.
- Bradley, G. (2006). *Nelson maths: Building mental strategies skillbook 3*. South Melbourne, VIC, Australia: Nightingale Press.
- Brown, G. T. L., & Farruggia, S. P. (2012). *Cleaning data: Practices that reduce the threat of garbage in, garbage out (GIGO)*. Unpublished manuscript. Faculty of Education, University of Auckland, Auckland, New Zealand.
- Burkhardt, H., & Bell, A. (2007). Problem solving in the United Kingdom. *ZDM*, 39(5-6), 395-403.
- Cambridge Education. (2016). *Teaching & learning quality in primary education: Assessment & recommendations (third primary education development program and post-PEDP 3 planning quality study)*. Retrieved from <https://www.dfat.gov.au/about-us/publications/Pages/bangladesh-teaching-learning-quality-primary-education-assessment-recommendations>
- Campbell, J. I., & Xue, Q. (2001). Cognitive arithmetic across cultures. *Journal of Experimental Psychology: General*, 130(2), 299.

- CAMPE. (2014). *Whither grade v examination?: An assessment of primary education completion examination in Bangladesh*. Dhaka, Bangladesh: Campaign for Popular Education (CAMPE).
- CAMPE. (2015). *Moving from MDG to SDG: Accelerate progress for quality primary education*. Dhaka, Bangladesh: Campaign for Popular Education (CAMPE).
- Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C.-P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal*, 26(4), 499-531.
- Carr, M., & Davis, H. (2001). Gender differences in arithmetic strategy use: A function of skill and preference. *Contemporary Educational Psychology*, 26(3), 330-347.
doi:10.1006/ceps.2000.1059
- Charles, R. I., & Lester, F. K. (1982). *Teaching problem solving: What, why and how?*. Palo Alto: Dale Seymour.
- Cheng, L. P. (2015). Error analysis for arithmetic word problems - A case study of primary three students in one Singapore school. *International Journal for Mathematics Teaching and Learning*. Retrieved from <https://www.cimt.org.uk/journal/lupien.pdf>
- Chowdhury, A. M. R., Nath, S. R., Choudhury, R. K., & Ahmed, M. (2002). *Renewed hope, daunting challenges: State of primary education in Bangladesh*. Dhaka: The University Press Limited.
- Clarke, D., Goos, M., & Morony, W. (2007). Problem solving and working mathematically: An Australian perspective. *ZDM*, 39(5-6), 475-490.
- Clements, M. K. (1980). Analyzing children's errors on written mathematical tasks. *Educational Studies in Mathematics*, 11(1), 1-21.
- Cockcroft Report. (1982). *Mathematics counts*. Retrieved from <http://www.educationengland.org.uk/documents/cockcroft/cockcroft1982.html>
- Codina, S. A., Cañadas, M. C., & Castro, E. (2015). Mathematical problem solving through sequential process analysis. *Education and Psychology*, 13(1), 73 - 110.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum Associates.
- Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). New York: Routledge.
- Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research*, 12(3), 229-238.
- Coquin-Viennot, D., & Moreau, S. (2003). Highlighting the role of the episodic situation model in the solving of arithmetical problems. *European Journal of Psychology of Education*, 18(3), 267-279.
- Creswell, J. W. (2012). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson Education.
- Creswell, J. W., & Clark, V. L. P. (2018). *Designing and conducting mixed methods research* (3rd ed.). Thousand Oaks, CA: SAGE publications.
- Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Thousand Oaks, CA: SAGE.
- Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry & research design: Choosing among five approaches* (4th ed.). Thousand Oaks, CA: SAGE.
- Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, Ohio: Cengage Learning.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443 - 503). Washington DC: American Council on Education.

- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, 58(4), 438-481.
- Deane, P., Sheehan, K. M., Sabatini, J., Futagi, Y., & Kostin, I. (2006). Differences in text structure and its implications for assessment of struggling readers. *Scientific Studies of Reading*, 10(3), 257 - 275.
- Denscombe, M. (2008). Communities of practice: A research paradigm for the mixed methods approach. *Journal of Mixed Methods Research*, 2(3), 270-283.
doi:10.1177/1558689808316807
- Denscombe, M. (2010). *Ground rules for social research: Guidelines for good practice* (2nd ed.). Berkshire, England: McGraw-Hill Education.
- Denscombe, M. (2014). *The good research guide: For small-scale social research projects* (5th ed.). Berkshire, England: McGraw-Hill Education.
- Department of Education. (2014). *Statuary guidance national curriculum in England: Mathematics programmes of study*. Retrieved from <https://www.gov.uk/government/publications/national-curriculum-in-england-mathematics-programmes-of-study/national-curriculum-in-england-mathematics-programmes-of-study>
- Dey, J., & Siddiquee, M. N.-E.-A. (2012). Assessing curriculum competencies: Review of Bangla test questions in primary education terminal examination. *Bangladesh Education Journal*, 11(2), 17 - 29.
- Directorate of Primary Education [DPE]. (2015). *প্রাথমিক শিক্ষা সমাপনী পরীক্ষা- ২০১৪ এর পরিসংখ্যান* [Primary education completion examination - 2014 statistics]. Retrieved from <https://dpe.portal.gov.bd/site/files/5f0c6429-3a4d-4dfa-97de-e5293d93103e>
- Directorate of Primary Education [DPE]. (2016). *প্রাথমিক শিক্ষা সমাপনী পরীক্ষা- ২০১৫ এর পরিসংখ্যান* [Primary education completion examination - 2015 statistics]. Retrieved from https://dpe.portal.gov.bd/sites/default/files/files/dpe.portal.gov.bd/page/e41dbfe5_f47a_4fcb_8352_049f5b6c36cc/PECE_2015_Statistics.pdf
- Directorate of Primary Education [DPE]. (2017a). *Annual primary school census 2017*. Dhaka: Ministry of Primary and Mass Education.
- Directorate of Primary Education [DPE]. (2017b). *Bangladesh primary education annual sector performance report - 2017*. Dhaka: Government of the People's Republic of Bangladesh.
- Dixon, D. D., & Worrell, F. C. (2016). Formative and summative assessment in the classroom. *Theory into Practice*, 55(2), 153-159. doi:10.1080/00405841.2016.1148989
- Dolin, J., Black, P., Harlen, W., & Tiberghien, A. (2018). Exploring relations between formative and summative assessment. In J. Dolin & R. Evans (Eds.), *Transforming assessment: Through an interplay between practice, research and policy* (pp. 53 - 80). Cham, Switzerland: Springer International Publishing.
- Dowling, P., & brown, A. (2010). *Doing research/ reading research: Re-interrogating education* (2nd ed.). New York: Routledge.
- Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3 - 25). London: Lawrence Erlbaum Associates Publishers.
- DPE. (2018). *The national student assessment 2017: Grades 3 and 5*. Dhaka, Bangladesh: Monitoring and Evaluation Division, Directorate of Primary Education (DPE).
- Duke, N. K., & Pearson, P. D. (2009). Effective practices for developing reading comprehension. *The Journal of Education*, 189(1/2), 107-122.
- Dunker, K. (1945). On problem-solving. *Psychological Monographs*, 58(5), i.

- Ebel, R. L. (1956). Obtaining and reporting evidence on content validity. *Educational and Psychological Measurement*, 16(3), 269-282.
doi:<https://doi.org/10.1177/001316445601600301>
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87(3), 215-251.
- Fennema, E., Carpenter, T. P., Jacobs, V. R., Franke, M. L., & Levi, L. W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher*, 27(5), 6 - 11.
- Fletcher, J. M. (2006). Measuring reading comprehension. *Scientific Studies of Reading*, 10(3), 323-330. doi:10.1207/s1532799xssr1003_7
- Fuson, K. C. (1992). Research on whole number addition and subtraction. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 243-295). New York: Macmillan.
- Fuson, K. C., Carroll, W. M., & Landis, J. (1996). Levels in conceptualizing and solving addition and subtraction compare word problems. *Cognition and Instruction*, 14(3), 345-371.
- Gadernann, A. M., Guhn, M., & Zumbo, B. D. (2011). Investigating the substantive aspect of construct validity for the satisfaction with life scale adapted for children: A focus on cognitive processes. *Social Indicators Research*, 100(1), 37-60. doi:10.1007/s11205-010-9603-x
- Gardner, J. (2010). Developing teacher assessments: An introduction. In J. Gardner, W. Harlen, L. Hayward, G. Stobart, & M. Montgomery (Eds.), *Developing teacher assessment* (pp. 1 - 11). New York: Open University Press.
- Gravemeijer, K., & Stephan, M. (2002). Emergent models as an instructional design heuristic. In K. Gravemeijer, R. Lehrer, B. V. Oers, & L. Verschaffel (Eds.), *Symbolizing, modeling and tool use in mathematics education* (pp. 145 - 169). Dordrecht: Kluwer Academic Publishers.
- Greiff, S., Holt, D. V., & Funke, J. (2013). Perspectives on problem solving in educational assessment: Analytical, interactive, and collaborative problem solving. *The Journal of Problem Solving*, 5(2), 71 - 91.
- Griffin, P., Francis, M., & Robertson, P. (2018). Guttman analysis. In P. Griffin (Ed.), *Assessment for teaching* (2nd ed., pp. 198 - 225). Cambridge, UK: Cambridge University Press.
- Habiba, U., Abedin, M. A., & Shaw, R. (2013). Disaster education in Bangladesh: Opportunities and challenges. In R. Shaw, F. Mallick, & A. Islam (Eds.), *Disaster risk reduction approaches in Bangladesh* (pp. 307 - 330). Tokyo: Springer.
- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65 - 110). Westport: Praeger Publishers.
- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Oxon, UK: Routledge.
- Halmos, P. (1980). The heart of mathematics. *American Mathematical Monthly*, 87, 519 - 524.
- Hambleton, R. K. (1993). *Translating achievement tests for use in cross-national studies* (ED358128). Retrieved from <https://files.eric.ed.gov/fulltext/ED358128.pdf>
- Harkness, J. A., & Schoua-Glusberg, A. (1998). Questionnaires in translation. *ZUMA-Nachrichten Spezial*, 7, 87 - 125.
- Hartnett, J. E. (2007). *Categorisation of mental computation strategies to support teaching and to encourage classroom dialogue*. Paper presented at the 30th annual conference of the Mathematics Education Research Group of Australasia - Mathematics: Essential Research, Essential Practice, Hobart, Tasmania.
- Hartnett, J. E. (2008). *Capturing students' thinking about strategies used to solve mental computation by giving students access to a pedagogical framework*. Paper presented at the Navigating Currents and Charting Directions: 31st Annual Conference of the Mathematics Education Research Group of Australasia, St Lucia, Brisbane.

- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81-112.
- Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, 58(1), 78-79. doi:10.1037/0003-066X.58.1.78
- Hossain, M. B. (2015). *Quality of mathematics teaching and learning in the primary schools of Bangladesh*. (MPhil), University of Dhaka, Dhaka. Retrieved from <http://repository.library.du.ac.bd:8080/handle/123456789/606>
- Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation. *Social Indicators Research*, 103(2), 219-230. doi:https://doi.org/10.1007/s11205-011-9843-4
- Hudson, R. F., Lane, H. B., & Pullen, P. C. (2005). Reading fluency assessment and instruction: What, why, and how? *The Reading Teacher*, 58(8), 702 - 714.
- Ineson, G. (2007). Year 6 children: Has the new British mathematics curriculum helped their mental computation? *Early Child Development and Care*, 177(5), 541-555. doi:10.1080/03004430600581895
- Institute of Educational Development [IED]. (2009). *Report on listening workshop for quality science and mathematics education in primary schools*. Dhaka: Institute of Educational Development, BRAC University.
- International Bureau of Education. (2011). *World data on education*. Retrieved from <http://www.ibe.unesco.org/en/document/world-data-education-seventh-edition-2010-11>
- Islam, M. S. (2016). Dilemma of the high-stake public examination for primary education in Bangladesh: Can decentralization help? . *Bangladesh Education Journal*, 15(2), 47 - 52.
- Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners. Technical Report 44*. Retrieved from <https://files.eric.ed.gov/fulltext/ED495909.pdf>
- Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development*, 48(4), 63-85.
- Kamei, T. (2019). *Thinking to learn: An investigation of assessing and teaching thinking skills for students with disability*. (Unpublished doctoral dissertation), The University of Melbourne, Melbourne, Australia,
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17 - 64). Westport: Praeger Publishers.
- Kaur, B., & Toh, T. L. (2011). Mathematical problem solving: Linking theory and practice. In O. Zaslavsky & P. Sullivan (Eds.), *Constructing knowledge for teaching secondary mathematics: Tasks to enhance prospective and practicing teacher learning* (pp. 177-188). New York: Springer.
- Kawasaki, M., Morita, T., Umezawa, M., & Koike, M. (2013). The relation between the formula for solution and the components of drawings for arithmetic word problem solving of the first-grade students in elementary school. *Educational Technology Research*, 36(1), 11-21.
- Kean, J., & Reilly, J. (2014a). Classical test theory. In F. M. Hammond, J. F. Malec, T. G. Nick, & R. M. Buschbacher (Eds.), *Handbook for clinical research: Design, statistics, and implementation* (pp. 208 - 210). New York: Demos Medical Publishing.
- Kean, J., & Reilly, J. (2014b). Item response theory. In F. M. Hammond, J. F. Malec, T. G. Nick, & R. M. Buschbacher (Eds.), *Handbook of clinical research: Design, statistics, and implementation* (pp. 211 - 214). New York: Demos Medical Publishing.
- Khan, A. F. M. K., Matin, S., Begum, H. B., & Uddin, D. M. M. (2013). *প্রাথমিক গণিত: তৃতীয় শ্রেণী* [Elementary mathematics: Class three]. Dhaka: National Curriculum and Textbook Board.

- Khan, N. J., Azim, F., & Ahmed, S. S. (2009). The practice of class work and practical work within SBA in secondary school classroom. *NAEM Journal*, 5(9), 15 - 24.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.
- Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review*, 92(1), 109.
- Kivunja, C., & Kuyini, A. B. (2017). Understanding and applying research paradigms in educational contexts. *International Journal of Higher Education*, 6(5), 26 - 41. doi:10.5430/ijhe.v6n5p26
- Kohler, W. (1925). *The mentality of apes*. New York: Liveright.
- Lamprianou, I., & Ahanasou, J. A. (2009). *A teacher's guide to educational assessment*. Boston: Sense Publishers.
- Leighton, J. P., & Gokiert, R. J. (2008). Identifying potential test item misalignment using student verbal reports. *Educational Assessment*, 13(4), 215-242. doi:10.1080/10627190802602384
- Leighton, J. P., Heffernan, C., Cor, M. K., Gokiert, R. J., & Cui, Y. (2011). An experimental test of student verbal reports and teacher evaluations as a source of validity evidence for test development. *Applied Measurement in Education*, 24(4), 324-348. doi:10.1080/08957347.2011.607061
- Lemaire, P. (2010). Cognitive strategy variations during aging. *Current Directions in Psychological Science*, 19(6), 363-369.
- Lemaire, P., & Arnaud, L. (2008). Young and older adults' strategies in complex arithmetic. *The American Journal of Psychology*, 121(1), 1-16.
- Lester, F. K., & Kehle, P. E. (2003). From problem solving to modeling: The evolution of thinking about research on complex mathematical activity. In R. Lesh & H. Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching* (pp. 501-518). New Jersey: Lawrence Erlbaum Associates.
- Lewis, A. B., & Mayer, R. E. (1987). Students' miscomprehension of relational statements in arithmetic word problems. *Journal of Educational Psychology*, 79(4), 363.
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: SAGE.
- Linninger, C., Kunina-Habenicht, O., Emmenlauer, S., Dicke, T., Schulze-Stocker, F., Leutner, D., Seidel, T., Terhart, E., & Kunter, M. (2015). Assessing teachers' educational knowledge: Construct specification and validation using mixed methods. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 47(2), 72 - 83.
- Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36(8), 437-448. doi:DOI: 10.3102/0013189X0731 1286
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3(3), 635-694. doi:https://doi.org/10.2466/pr0.1957.3.3.635
- Lukin, L. E., Bandalos, D. L., Eckout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice*, 23(2), 26-32.
- Lutfuzzaman, A. A. K. M., Muhammad, N. M., & Hasan, A. S. M. T. (2006). Developing a quality mathematics education culture in Bangladesh. *Bangladesh Education Journal*, 5(2), 25 - 34.
- Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, 110(1), 27-45. doi:http://dx.doi.org/10.1037/edu0000205
- Martin, M. O., Mullis, I. V. S., & Chrostowski, S. J. (Eds.). (2004). *TIMSS 2003 technical report: Findings from IEA's trends in international mathematics and science study at the*

- fourth and eighth grades*. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
- Mayer, R. (1982). The psychology of mathematical problem solving. In F. K. Lester & J. Garofalo (Eds.), *Mathematical problem solving: Issues in research* (pp. 1-13). Philadelphia: Franklin Institute Press.
- Mayer, R. (1995). The search for insight: Grappling with Gestalt psychology's unanswered questions. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 3 - 32). Cambridge, MA: MIT Press.
- Mayer, R. (2013). Problem solving. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 769 - 768). New York: Oxford University Press.
- Mayer, R., & Wittrock, M. (2009, 23 December 2009). Problem solving. Retrieved from www.education.com/reference/article/problem-solving1
- Mayer, R., & Wittrock, M. C. (2006). Problem solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 287 - 304). Mahwah, NJ: Erlbaum.
- McIntosh, A. J. (2002). *Developing informal written computation*. Paper presented at the Annual Conference of the Australian Association for Research in Education, University of Queensland, Brisbane.
- McKay, R. B., Breslow, M. J., Sangster, R. L., Gabbard, S. M., Reynolds, R. W., Nakamoto, J. M., & Tarnai, J. (1996). Translating survey questionnaires: Lessons learned. *New Directions for Evaluation*, 1996(70), 93-104.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13 - 103). New York: American Council on Education; Macmillan Publishing Company.
- Miller, K. (2014). Introduction. In K. Miller, V. Chepp, S. Willson, & J. L. Padilla (Eds.), *Cognitive interviewing methodology* (pp. 1 - 6). Hoboken, New Jersey: John Wiley & Sons.
- Miller, K., Willson, S., Chepp, V., & Ryan, M. J. (2014). Analysis. In K. Miller, V. Chepp, S. Willson, & J. L. Padilla (Eds.), *Cognitive interviewing methodology* (pp. 35 - 50). Hoboken, New Jersey: John Wiley & Sons.
- Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching*. New Jersey: Pearson Education.
- Mills, G. E., & Gay, L. R. (2016). *Educational research: Competencies for analysis and applications* (11th ed.). Essex, England: Pearson Education Limited.
- Ministry of Education. (2010). *National education policy 2010*. Dhaka: Ministry of Education, Government of the People's Republic of Bangladesh.
- Ministry of Primary and Mass Education [MoPME]. (2014). *প্রাথমিক শিক্ষায় পাঁচ বছরের অর্জন* [Achievement in primary education in five years]. Retrieved from [https://mopme.portal.gov.bd/sites/default/files/files/mopme.portal.gov.bd/page/29a0eac0_d74d_46a5_91df_2074d4c9ddca/AR_2014_Complete_small%20size%20\(1\).pdf](https://mopme.portal.gov.bd/sites/default/files/files/mopme.portal.gov.bd/page/29a0eac0_d74d_46a5_91df_2074d4c9ddca/AR_2014_Complete_small%20size%20(1).pdf)
- Ministry of Primary and Mass Education [MoPME]. (2015). *Education for all 2015 national review*. Dhaka, Bangladesh: Ministry of Primary and Mass Education (MoPME), Government of Bangladesh.
- Mollah, S. H., Ullah, A. M. M. A., Halder, D. A., & Dhali, S. K. (2013a). *Elementary mathematics: Class five*. Dhaka: National Curriculum and Textbook Board.
- Mollah, S. H., Ullah, A. M. M. A., Halder, D. A., & Dhali, S. K. (2013b). *Elementary mathematics: Class four*. Dhaka: National Curriculum and Textbook Board.
- Mollah, S. H., Ullah, A. M. M. A., Halder, D. A., & Dhali, S. K. (2013c). *প্রাথমিক গণিত: চতুর্থ শ্রেণী* [Elementary mathematics: Class four]. Dhaka: National Curriculum and Textbook Board.
- Mollah, S. H., Ullah, A. M. M. A., Halder, D. A., & Dhali, S. K. (2013d). *প্রাথমিক গণিত: পঞ্চম শ্রেণী* [Elementary mathematics: Class five]. Dhaka: National Curriculum and Textbook Board.

- Mullis, I. V. S., & Martin, M. O. (Eds.). (2013). *TIMSS 2015 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education Boston College, and International Association for the Evaluation of Educational Achievements (IEA).
- NAPE. (2015). *Review of the results of the 35% competency based test items in the 2014 grade 5 primary completion examination*. Mymensingh, Bangladesh: National Academy for Primary Education (NAPE).
- NAPE. (2018). *Effectiveness of lesson study: A case of enhancing quality of mathematics teaching learning in the primary education of Bangladesh*. Mymensingh, Bangladesh: National Academy for Primary Education (NAPE).
- National Council of Teachers of Mathematics [NCTM]. (1989). *Principles and standards for school mathematics*. Reston, VA: NCTM.
- National Council of Teachers of Mathematics [NCTM]. (2000). *Principles and standards for school mathematics*. Reston, VA: The National Council of Teachers of Mathematics.
- National Curriculum and Textbook Board [NCTB]. (2015a). *প্রাথমিক গণিত: চতুর্থ শ্রেণি [Elementary mathematics: Class four]*. Dhaka, Bangladesh: National Curriculum and Textbook Board.
- National Curriculum and Textbook Board [NCTB]. (2015b). *প্রাথমিক গণিত: তৃতীয় শ্রেণি [Elementary mathematics: Class three]*. Dhaka: National Curriculum and Textbook Board (NCTB).
- National Curriculum and Textbook Board [NCTM]. (2011). *Primary curriculum (English version)*. Dhaka: National Curriculum and Textbook Board (NCTB).
- National Research Council. (2001). *Classroom assessment and the national science education standards*. Washington DC: National Academy Press.
- NCTB. (2009). *Elementary mathematics: Class three*. Dhaka: National Curriculum and Textbook Board (NCTB).
- NCTB. (2012). *Primary curriculum 2011: Class 1 - 5*. Dhaka, Bangladesh: National Curriculum and Textbook Board (NCTB), Government of Bangladesh.
- NCTB. (2016a). *শিক্ষক সংস্করণ প্রাথমিক গণিত চতুর্থ শ্রেণি [Teacher edition primary mathematics grade four]*. Dhaka, Bangladesh: National Curriculum and Textbook Board (NCTB).
- NCTB. (2016b). *শিক্ষক সংস্করণ প্রাথমিক গণিত তৃতীয় শ্রেণি [Teacher edition primary mathematics grade three]*. Dhaka, Bangladesh: National Curriculum and Textbook Board (NCTB).
- NCTB. (2016c). *শিক্ষক সংস্করণ প্রাথমিক গণিত পঞ্চম শ্রেণি [Teacher edition primary mathematics grade five]*. Dhaka, Bangladesh: National Curriculum and Textbook Board (NCTB).
- Neumann, I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test. *International Journal of Science Education*, 33(10), 1373-1405.
doi:<https://doi.org/10.1080/09500693.2010.511297>
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. New Jersey: Prentice Hall.
- Nickerson, R. S. (1999). Enhancing creativity. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 390 - 428). New York: Cambridge University Press.
- Nielsen, J. (1994). Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies*, 41(3), 385-397.
- Nolen, S. B. (2011). The role of education systems in the link between formative assessment and motivation. *Theory into Practice*, 50(4), 319-326.
- OECD. (2010). *PISA 2012 field trial problem solving framework (draft subject to possible revision after the field trial)*. Retrieved from https://moam.info/pisa-2012-problem-solving-framework_59828b641723ddf156c8be41.html

- OECD. (2012). *PISA 2012 Assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Retrieved from <http://dx.doi.org/10.1787/9789264190511-en>
- OECD. (2014). *PISA 2012 results: Creative problem solving (volume v): Students' skills in tackling real-life problems*. Retrieved from <http://dx.doi.org/10.1787/9789264208070-en>
- OECD. (2020). *PISA for development technical report*. Retrieved from <https://www.oecd.org/pisa/pisa-for-development/pisafordevelopment2018technicalreport/>
- Opel, A., Zaman, S. S., Khanom, F., & Aboud, F. E. (2012). Evaluation of a mathematics program for preprimary children in rural Bangladesh. *International Journal of Educational Development*, 32, 104 - 110. doi:10.1016/j.ijedudev.2011.01.013
- Osteen, P. (2010). An introduction to using multidimensional item response theory to assess latent factor structures. *Journal of the Society for Social Work and Research*, 1(2), 66-82. doi:10.5243/jsswr.2010.6
- Österholm, M. (2006). Characterizing reading comprehension of mathematical texts. *Educational Studies in Mathematics*, 63, 325-346. doi:10.1007/s10649-005-9016-y
- Padilla, J. L., & Benitez, I. (2014). Validity evidence based on response processes. *Psicothema*, 26(1), 136-144. doi:10.7334/psicothema2013.259
- Phongphanngam, S., & Lach, H. W. (2019). Cross-cultural instrument translation and adaptation: Challenges and strategies. *Pacific Rim International Journal of Nursing Research*, 23(2), 170-179.
- Polya, G. (1973). *How to solve it: A new aspect of mathematical method*. Princeton, NJ: Princeton University Press.
- Pongsakdi, N., Kajamies, A., Veermans, K., Lertola, K., Vauras, M., & Lehtinen, E. (2020). What makes mathematical problem word problem solving challenging? Exploring the roles of word problem characteristics, text comprehension, and arithmetic skills. *ZDM*, 52(1), 33-44. doi:<https://doi.org/10.1007/s11858-019-01118-9>
- Pouzevara, S. L., & Khan, R. (2007). *Learning communities enabled by mobile technology: A case study of school-based, in-service secondary teacher training in rural Bangladesh*. Retrieved from <https://shared.rti.org/content/learning-communities-enabled-mobile-technology-case-study-school-based-service-secondary>
- Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for testing and evaluating survey questions. In S. Presser, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, J. M. Rothgeb, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 1 - 22). Hoboken, New Jersey: John Wiley & Sons.
- Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using stata* (3rd ed.). College Station, Texas: Stata Press.
- Rahman, M. (2010). Classroom assessment and student learning: An exploration of secondary school teacher practices. *NAEM Journal*, 5(10), 14-25.
- Reeves, T. D., & Marbach-Ad, G. (2016). Contemporary test validity in theory and practice: A primer for discipline-based education researchers. *CBE-Life Sciences Education*, 15(rm1), 1-15. doi:DOI:10.1187/cbe.15-08-0183
- Richards, J., & Islam, M. S. (2018). Assessing literacy and numeracy among primary school students: A pilot survey in rural Bangladesh. *International Journal of Educational Development*, 61, 55 - 63. doi:10.1016/j.ijedudev.2017.12.001
- Richards, J., & Vining, A. R. (2016). Achieving better primary school outcomes in a context of weak education governance: A case study of Bangladesh. *Bangladesh Education Journal*, 15(2), 9 - 25

- Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, 135(2), 207-236. doi:10.1037/0096-3445.135.2.207
- Riley, M. S., Greeno, J. G., & Heller, J. I. (1984). Development of children's problem-solving ability in arithmetic. In H. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 153-196). Orlando, FL: Academic Press
- Robertson, S. I. (2001). *Problem solving*. Philadelphia, PA: Psychological Press.
- Roy, G. (2016). *Competency-based assessment in primary education in Bangladesh: A review (SHARE education monograph 6)*. Dhaka: SHARE TA office.
- Sawada, Y., Mahmud, M., Seki, M., Le, A., & Kawarazaki, H. (2017). *Individualized self-learning program to improve primary education: Evidence from a randomized field experiment in Bangladesh (working paper no. 156)*. Retrieved from https://www.jica.go.jp/jica-ri/publication/workingpaper/wp_156.html
- Schoenfeld, A. H. (1985). *Mathematical problem solving*. Orlando, FL: Academic Press.
- Schoenfeld, A. H. (2013). Reflections on problem solving theory and practice. *The Mathematics Enthusiast*, 10(1), 9-34.
- Schroeder, T. L., & Lester, F. K. (1989). Developing understanding in mathematics via problem solving. In P. Trafton & A. Shulte (Eds.), *New directions for elementary school mathematics (1989 yearbook)* (pp. 31-42). Reston, VA: NCTM.
- Scotland, J. (2012). Exploring the philosophical underpinnings of research: Relating ontology and epistemology to the methodology and methods of the scientific, interpretive, and critical research paradigms. *English Language Teaching*, 5(9), 9-16. doi:10.5539/elt.v5n9p9
- Sheatsley, P. (1983). Questionnaire construction and item writing. In P. Rossi, J. Wright, & A. Anderson (Eds.), *Handbook of survey research* (pp. 195 - 230). San Diego, CA: Academic Press.
- Shekh, M. N. I. (2006). *The state of primary education quality in Bangladesh: An evaluation*. Paper presented at the NETREED conference 2005, Beitostølen, Norway.
- Shohel, M. M. C., & Banks, F. (2010). *Teachers' professional development through the English in action secondary teaching and learning programme in Bangladesh: Experience from the UCEP schools*. Paper presented at the 2nd World Conference on Educational Sciences, Istanbul, Turkey.
- Siegler, R. S. (1999). Strategic development. *Trends in Cognitive Sciences*, 3(11), 430-435.
- Sireci, S. G. (2009). Packing and unpacking sources of validity evidence. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19-37). Charlotte, NC: Information Age Publishing.
- Stacey, K. (2003). The need to increase attention to mathematical reasoning. In J. Lokan, B. McRae, & H. Hollingsworth (Eds.), *Teaching mathematics in Australia: Results from the TIMSS 1999 video study. TIMSS Australia Monograph No. 5*. (pp. 119 - 126). Camberwell, VIC: Australian Council for Educational Research.
- Stacey, K. (2005). The place of problem solving in contemporary mathematics curriculum documents. *Journal of Mathematical Behavior*, 24(3), 341-350.
- Stacey, K., Price, B., Gvozdenko, E., & Steinle, V. (2013, 15/03/2017). Specific mathematics assessments that reveal thinking. Retrieved from <http://www.smartvic.com/teacher/>
- Stacey, K., & Turner, R. (2015). *Assessing mathematical literacy: The PISA experience*. New York: Springer.
- Strickland, J. (2018). *The design and use of a learning progression of emergent numeracy to support the instruction of students with intellectual and developmental disability*. (Unpublished doctoral dissertation), The University of Melbourne, Melbourne, Australia,

- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257-285.
- Sweller, J., Clark, R., & Kirschner, P. (2010). Teaching general problem-solving skills is not a substitute for, or a viable addition to, teaching mathematics. *Notices of the American Mathematical Society*, 57(10), 1303-1304.
- Tai, W.-C., & Lin, S.-W. (2015). Relationship between problem-solving style and mathematical literacy. *Educational Research and Reviews*, 10(11), 1480-1486.
- Tambara, C. T. (2015). *Unpacking teachers' pedagogical content knowledge and skills to develop learners' problem solving skills in mathematics*. (Doctoral dissertation), Stellenbosch University, South Africa, Retrieved from http://scholar.sun.ac.za/bitstream/handle/10019.1/96772/tambara_unpacking_2015.pdf?sequence=1&isAllowed=y
- Thevenot, C., & Barrouillet, P. (2015). Arithmetic word problem solving and mental representations. In R. C. Kadosh & A. Dowker (Eds.), *The Oxford handbook of numerical cognition* (pp. 158 - 179). New York: Oxford University Press.
- Thevenot, C., Barrouillet, P., & Fayol, M. (2004). Mental representation and procedures in arithmetic word problems: The effect of the position of the question. *Année Psychologique*, 104(4), 683-699.
- Thevenot, C., Devidal, M., Barrouillet, P., & Fayol, M. (2007). Why does placing the question before an arithmetic word problem improve performance? A situation model account. *The Quarterly Journal of Experimental Psychology*, 60(1), 43-56.
- Thorndike, E. L. (1911). *Animal intelligence*. New York: The Macmillan Company.
- Tuli, F. (2010). The basis of distinction between qualitative and quantitative research in social science: Reflection on ontological, epistemological and methodological perspectives. *Ethiopian Journal of Education and Sciences*, 6(1), 97-108.
- Tuma, D. T., & Reif, F. (1980). *Problem solving and education: Issues in teaching and research*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Uddin, M. M. (2005). *Numeracy in Bangladesh: Terminal competencies of mathematics at primary education*. Paper presented at the Second International Education Cooperation Symposium, Tokyo. <http://www.cried.tsukuba.ac.jp/math/sym/mohsin0123.pdf>
- UNESCO. (2006). *Teachers and education quality: Monitoring global needs for 2015*. Montreal: UNESCO Institute for Statistics.
- UNICEF. (2009). *Quality primary education in Bangladesh*. Dhaka: UNICEF Bangladesh.
- Vergnaud, G. (1982). A classification of cognitive tasks and operations of thought involved in addition and subtraction problems. In T. P. Carpenter, J. M. Moser, & T. A. Romberg (Eds.), *Addition and subtraction: A cognitive perspective* (pp. 39 - 59). Hillsdale, NJ: Lawrence Erlbaum.
- Verschaffel, L., & De Corte, E. (1997). Word problems: A vehicle for promoting authentic mathematical understanding and problem solving in the primary school? . In T. Nunes & P. Bryant (Eds.), *Learning and teaching mathematics: An international perspective* (pp. 69 - 97). Hove, UK: Psychology Press.
- Weaver, K. (2018). Pragmatic paradigm. In B. B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation* (pp. 1286 - 1288). London: SAGE.
- Willatts, P. (1990). Development of problem solving strategies in infancy. In D. F. Bjorklund (Ed.), *Children's strategies: Contemporary views of cognitive development* (pp. 23-66). Hillsdale, NJ: Lawrence Erlbaum.
- Willis, G. (2004). Cognitive interviewing revisited: A useful technique, in theory? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 23-43). Hoboken, New Jersey: John Wiley & Sons.

- Willis, G. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: SAGE.
- Willis, G. (2015). *Analysis of the cognitive interview in questionnaire design*. New York: Oxford University Press.
- Willis, G., & Fuson, K. C. (1988). Teaching children to use semantic drawings to solve addition and subtraction word problems. *Journal of Educational Psychology*, 80(2), 192 - 201.
- Wilson, M. (2002). Assessment tools: Psychometric and statistical. In J. W. Guthrie (Ed.), *Encyclopedia of education* (2nd ed., pp. 139 - 145). New York: Macmillan Reference.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. New Jersey: Lawrence Erlbaum Associates.
- Wolfe, E. W., & Smith, E. V., Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: Part I - Instrument development tools. In E. V. J. Smith & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 202 - 242). Maple Grove, MN: JAM Press.
- Wong, T. T.-Y., & Ho, C. S.-H. (2017). Component processes in arithmetic word-problem solving and their correlates. *Journal of Educational Psychology*, 109(4), 520-531. doi:<http://dx.doi.org/10.1037/edu0000149>
- World Bank. (2011). *Bangladesh - third primary education development program (PEDP III)*. Retrieved from <http://documents.worldbank.org/curated/en/407251468209950682/Bangladesh-Third-Primary-Education-Development-Program-PEDPIII>
- World Bank. (2013). *Bangladesh education sector review: Seeding fertile grounds - education that works for Bangladesh*. Retrieved from <https://openknowledge.worldbank.org/handle/10986/17853>
- World Bank. (2016). *Bangladesh - country snapshot*. Retrieved from <http://documents.worldbank.org/curated/en/654391476782032287/Bangladesh-Country-snapshot>
- World Bank. (2018a). *International development association program appraisal document: Quality learning for all program for fourth primary education development program*. Retrieved from <http://documents1.worldbank.org/curated/en/857071529206219039/text/BANGLADESH-QLEAP-PAD-05252018.txt>
- World Bank. (2018b). *World development report 2018: Learning to realize education's promise*. Retrieved from <https://www.worldbank.org/en/publication/wdr2018>
- Wright, R. J., Martland, J., & Stafford, A. K. (2006). *Early numeracy* (2nd ed.). Thousand Oaks, CA: SAGE.
- Wu, H.-H. (2016). *Teaching school mathematics: Algebra*. Providence, Rhode Island: American Mathematical Society.
- Wu, M. (2003). *The application of item response theory to measure problem-solving proficiencies* (Doctoral dissertation), The University of Melbourne.
- Wu, M., & Adams, R. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics Education Research Journal*, 18(2), 93-113.
- Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, 14(4), 339 - 355.
- Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers: Theory into practice*. Singapore: Springer.
- Zhang, Z., & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education*, 16(4), 323-342.
- Zia-us-Sabur, & Ahmed, M. (2010). Multiple providers and access to primary education: The case of Bangladesh. *Prospects*, 40, 393 - 415. doi:10.1007/s11125-010-9161-7

- Zoanetti, N., Griffin, P., & Adams, R. (2006). *Applications of item response theory to identify and account for suspect rater data*. Paper presented at the AARE Conference, Adelaide. <http://www.aare.edu.au/06pap/zoa06310.pdf>
- Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 65-82). Charlotte, NC: Information Age Publishing.

Appendices

Appendix 1.1: Assessment of word problem solving in the Teacher Editions (grades three to five)

Teacher edition primary mathematics grade three (NCTB, 2016b)
The second chapter of this book discusses addition addressing grade three learning outcomes 9.1 (Will be able to add a maximum of four numbers of two, three or four digits without carrying) and 9.2 (Will be able to add a maximum of four numbers of two, three or four digits with carrying (the sum will not exceed 10000)). There are some word problems in this chapter but the instructions for teachers focus on the operation (addition) only. Observations, written and oral assessments are mentioned as assessment techniques and teachers are instructed to determine whether students were able to successfully do the addition. There is no mention of checking any other aspects of the problem solving. [P – 33 to 52] The first few sections of chapter three [P – 53 to] of the book discusses subtraction addressing grade three learning outcomes 10.1 (will be able to subtract a three- or four-digit number from a three or four digit number without carrying), 10.2 (will be able to subtract from a number of three or four digits a smaller number of three or four digits with carrying), and 10.3 (will be able to say and identify in subtraction the quantity which is subtracted from, from which another quantity is subtracted, the quantity which is subtracted and the result of subtraction). Similar to the addition chapter, there are some subtraction word problems in this chapter but the instructions focus on the operation. Suggested assessment techniques/ methods are same (observations, written and oral assessment) through which teachers are instructed to determine whether students were able to do the subtraction. For addition and subtraction problems in the abovementioned sections, teachers are simply instructed to focus on whether the students ‘were able to solve the problem’ that is, it is considered as a correct/incorrect, able/unable situation assessment. Moreover, it seems the Teacher Edition (teacher guide) instructs teachers to focus on the operation exclusively. For example, on page 58 of the TG, under assessment, teachers are instructed to use observation, oral and written responses and determine if students were able to use subtraction for four-digit numbers without borrowing. The final section of chapter three [P – 77 to 78] focuses on grade three learning outcome 11.1 (will be able to solve two-step problems involving addition and subtraction using numbers not exceeding four digits). There are several (two-step) addition and subtraction word problems in this section. Teachers are instructed to use observation and oral question as assessment methods/techniques while teaching these. The instruction for them is to focus on whether students were able to understand the two-stage addition/subtraction word problems and solve them. Chapter six of this book [P - 132] discusses grade three learning outcome 14.2 (will be able to solve problems using either addition or subtraction and using either

division or multiplication (in all the stages of the activity numbers having not more than two digits will be used but in case of division the divisor will be an one-digit number). However, there is no word problem discussed that focuses on addition and subtraction individually or together. All the word problems discussed have either multiplication or division within their combination.

The next chapter (chapter seven) includes grade learning outcome 15.2 (will be able to exchange coins and currency notes; be able to solve easy problems relating to day to day transaction involving coins and currency notes) [P – 137 to 148]. For assessing students' ability to solve problems involving monetary transactions, no assessment technique/strategy is mentioned. Teachers are instructed to assess whether students can convert different monetary denominations and do addition/subtraction. The instructions focus on whether students can or cannot do this and do not focus on any other skills (like real life knowledge, ability to understand the problem from reading the stem, etc.) [P - 146]

Chapter nine includes contents about measurement of length, weight and time. There are some word problems related to these areas here. However, the relevant learning outcome for word problem solving and such measurements is not part of grade three curriculum. That comes in grade five. Students are doing the word problems here, but no assessment technique/strategy is mentioned for word problems involving different units of measuring length, weight or time. Assessment instruction for teachers focuses on determining whether students can or cannot do addition/subtraction while converting different units and do not focus on any other skills (like real life knowledge, ability to understand the problem from reading the stem, etc.) [e.g., P - 178]

Teacher edition primary mathematics grade four (NCTB, 2016a)

There is not much focus on word problem solving in grade four mathematics book/teacher edition. Addition and subtraction are discussed in the second chapter of this book and chapter five discusses some word problems involving addition, subtraction, multiplication and division.

Even though there are some addition and subtraction word problems discussed in the second chapter, the focus of the teacher guide is on the operation and not the word problem solving. For example, while discussing addition in the teacher edition (teacher guide) oral or written assessment, assignment and observation are suggested as assessment methods to determine whether students were able to do addition of not more than five-digit numbers. Teachers are instructed to determine if the students were able to do the addition correctly. There is no mention of any assessment method/strategy to check if the students were able to understand the problem. [P – 38]

However, within the teaching learning activities [P – 38], there is a suggestion that teachers should be conscious that students are able to select the appropriate

mathematical sentence based on the word problem solving item.

Similar instructions/suggestions are given for subtraction as well.

The later part of this chapter discusses the grade four learning outcome 11.1 – ‘will be able to solve three-step problems involving addition and subtraction using numbers with no more than five-digits’. This outcome is further parsed here as this section starts with two-step problems and then moves on to three steps.

Observation, oral and written assessment are mentioned as assessment methods for these. As per the instruction here, assessment should focus on determining whether students were able to write a single mathematical sentence and solve the problem. There is no mention of specifically checking if students had difficulty in reading and understanding the problem, with calculation or communicating the results.

Chapter five addresses the grade four learning outcome 14.2 (Will be able to solve problems of three steps by using two or three of the methods of adding, subtracting, multiplying and dividing in the whole process, the numbers used will be of no more than three digits). There are a number of word problems in this chapter but similar to the second chapter, focus of the assessment discussed here is on the operations rather than the word problem solving. Observation, oral question, written response and assignment are mentioned as assessment methods/strategies here and teachers are instructed to determine if students understand the use of parentheses and sequence of different operations (addition, subtraction, multiplication and division while solving a problem. [P - 115 to 130]

Chapter ten of the grade four textbook and the teacher edition (teachers’ guide) on measurement. Conversion of different measurement units of length, weight, liquid is discussed here. There is some word problems that focus on conversion of units but there is none on addition/subtraction of different measurements. Observation, written and oral questions are mentioned as assessment techniques and teachers are instructed to focus on the conversion of units.

The next chapter (chapter eleven) discusses ‘time’. This chapter includes the grade four learning outcome 25.3 (will be able to carry out time related addition and subtraction and make use of them. There are some lessons focusing on word problems related to addition and subtraction of time. Assignments and oral questions are mentioned as assessment techniques/methods for these and teachers are instructed to determine if students were able to convert different units of time, if they were able to answer the questions. There is no mention of checking if students were able to understand the problem or focusing on any other difficulties they might have including understanding real life situations. [P – 267 - 274]

Teacher edition primary mathematics grade five (NCTB, 2016c)

Grade five mathematics textbook and teacher edition (teachers’ guide) has a chapter (chapter 3) that covers mathematical problems involving the four

operations (addition, subtraction, multiplication, and division). This chapter includes the grade five learning outcomes 14.1 (will be able to solve problems involving summing and subtracting) and 14.3 (will be able to solve mathematical problems using addition, subtraction, multiplication & division and use of brackets) [P – 21 to 38]. However, the emphasis is on students' learning how to use multiple operations together to solve a problem. Word problems are discussed but focus is on the operations rather than on how to get to the operations from the word problems. Teachers are instructed to use various assessment techniques/methods such as observation, assignment, oral and written questions but the instructions predominantly focuses on utilizing parenthesis for solving equations.

Chapter eleven discusses measurement [P – 227 to 262] of length, weight, volume and area. This chapter addresses grade five learning outcome 24.1 (will be able to perform the operations of addition, subtraction, multiplication and division related to the measurements of length and weight by using units in the metric system). Teachers are instructed to use observation, assignment, and oral questions as assessment methods/techniques but the assessment is supposed to determine whether students are able to convert the different measurement units. There is no discussion or instruction on how to assess students' ability of solving the word problems.

Appendix 3.1: Construct map and item content specifications for the pilot study items

Separate construct maps were developed for the five strands of the hypothesized theoretical framework of Arithmetic Word Problem Solving (AWPS). These are shown below.

Comprehension of the problem:

Construct map and items for strand 1: Comprehension of the problem				
Ability level	Content specification	Test Item	Item Difficulty level	
Able to understand information sequence in a short sentence	Decoding information from a sequence of numbers (short and simple sentence).	Item 1	Low	
Able to understand information sequence in a long sentence	Decoding information from a sequence (Long and complex sentence)	Item 7	Low	
Able to discern abstract comparison of information	Comparison and Inference from abstract text	Item 13	Medium	
Able to compare different units by identifying relevant information within the text	Comparison and inference from quantities given in different units (Long and complex sentence)	Item 19	Medium	
Able to compare directions by identifying relevant information within the text	Comparison and inference from concepts of direction given in text	Item 25	High	
Able to compare abstract quantities given in a complex sentence	Comparison of abstract quantities and inference from long text	Item 31	High	

Mapping conceptual relations to a numerical representation:

Construct map and items for strand 2: Mapping conceptual relations to a numerical representation			
Ability level	Content specification	Test Item	Item Difficulty level
Able to numerically map a small 'change: result unknown' subtraction word problem	Mapping a subtraction word problem with two numbers	Item 2	Low
Able to numerically map a small 'change: result unknown' addition word problem	Mapping an addition word problem with two numbers	Item 8	Low
Able to numerically map a large 'combine: value unknown' addition word problem	Mapping an addition word problem with three numbers where the combined value is unknown	Item 14	Medium
Able to numerically map a small 'change: start unknown' subtraction word problem	Mapping a subtraction word problem with two numbers where the starting value is unknown	Item 20	Medium
Able to numerically map a complex 'combine: combine value unknown' addition word problem	Mapping an addition word problem with three numbers where numbers have to be combined in two steps	Item 26	High
Able to numerically map a complex 'compare: compared quality unknown' subtraction word problem	Mapping a subtraction word problem with three numbers where two steps are necessary to find the compared quality	Item 32	High

Selecting an appropriate strategy for solution:

Construct map and items for strand 3: Selecting an appropriate strategy for solution			
Ability level	Content specification	Test Item	Item Difficulty level
Able to add two small numbers without carrying	Addition of two two-digit numbers without carrying	Item 3	Low
Able to subtract two small numbers without borrowing	Subtracting two two-digit numbers without borrowing	Item 9	Low
Able to add two small numbers with carrying	Addition of two two-digit numbers using carrying once	Item 15	Medium
Able to subtract two small numbers with borrowing	Subtracting two two-digit numbers with borrowing once	Item 21	Medium
Able to do large addition with carrying	Addition of three numbers (four digit + three digit+ two digit) with carrying once	Item 27	High
Able to do large subtraction that require addition	Subtracting three numbers (four digits – three digits – two digits), without borrowing (adding two numbers first to subtract)	Item 33	High
Able to choose a suitable mental-arithmetic strategy for a small addition	Addition of two two-digit numbers	Item 39	Medium
Able to choose a suitable mental arithmetic strategy for a small subtraction	Addition of two numbers (two digits – one digit)	Item 40	Medium

Performing the calculations on the stored numbers:

Construct map and items for strand 4: Performing the calculations on the stored numbers			
Ability level	Content specification	Test Item	Item Difficulty level
Able to add two small numbers	Addition of two numbers (maximum two digits)	Item 4	Low
Able to subtract small simple numbers	Subtracting two numbers (maximum two-digits)	Item 10	Low
Able to add two small numbers approximating for the result	Addition of two numbers (maximum two digits) and approximation for result	Item 16	Medium
Able to subtract two small numbers approximating for the result	Subtracting two numbers (maximum two digits) and approximation for result,	Item 22	Medium
Able to add two large numbers approximating for result	Addition of two numbers and approximation for result, maximum four digits	Item 28	High
Able to subtract two large numbers	Subtracting two numbers, maximum four digits	Item 34	High

Verifying the solution and communicating the results:

Construct map and items for strand 5: Verifying the solution and communicating the results			
Ability level	Content specification	Test Item	Item Difficulty level
Able to verify solution and result of a small addition	Verifying the solution and communicating the result for adding two numbers (maximum three-digit number)	Item 5	Low
Able to verify solution and result of a small subtraction	Verifying the solution and communicating the result for subtracting two numbers (maximum three-digit number)	Item 11	Low
Able to verify solution and result for a large addition	Verifying the solution and communicating the result for adding three numbers (maximum four-digit number)	Item 17	Medium
Able to verify solution and result for a large subtraction	Verifying the solution and communicating the result for subtracting two numbers (maximum four-digit number)	Item 23	Medium
Able to verify solution and result for a large addition while applying real-life understanding	Verifying the solution and communicating the result for an addition with understanding of real-life situation	Item 29	High
Able to verify solution and result for a small subtraction while applying real-life understanding	Verifying the solution and communicating the result for a subtraction with understanding of real-life situation	Item 35	High
Able to verify solution and result with knowledge of time-format expression	Verifying the solution and the result based on correct format for expressing time	Item 37	High
Able to verify solution and result with knowledge of unit of currency	Verifying the solution and the result based on correct format for unit of currency	Item 38	High

Appendix 3.2: Construct map and item content specifications for the final test items

Separate construct maps were developed for the five strands of the hypothesized theoretical framework of Arithmetic Word Problem Solving (AWPS). Based on the findings of the pilot study, the construct map and item specification were adjusted for the final study items. These are shown below.

Comprehension of the problem:

Construct map and items for strand 1: Comprehension of the problem			
Ability level	Content specification	Test Item Number	Item Difficulty level
Able to understand information sequence in a small sentence	Decoding information from a sequence of numbers (short and simple sentence).	Item 1	Low
Able to understand information sequence in a long sentence	Decoding information from a sequence (Long and complex sentence)	Item 7	Low
Able to discern abstract comparison of information	Comparison and Inference from abstract text	Item 13	Medium
Able to compare different units by identifying relevant information within the text	Comparison and inference from quantities given in different units (Long and complex sentence)	Item 19	Medium
Able to compare directions by identifying relevant information within the text	Comparison and inference from concepts of direction given in text	Item 25	High
Able to compare abstract quantities given in a complex sentence	Comparison of abstract quantities and inference from long text	Item 31	High

Mapping conceptual relations to a numerical representation:

Construct map and items for strand 2: Mapping conceptual relations to a numerical representation			
Ability level	Content specification	Test Item Number	Item Difficulty level
Able to numerically map a small 'change: result unknown' subtraction word problem	Mapping a subtraction word problem with two numbers	Item 2	Low
Able to numerically map a small 'change: result unknown' addition word problem	Mapping an addition word problem with two numbers	Item 8	Low
Able to numerically map a large 'combine: value unknown' addition word problem	Mapping an addition word problem with three numbers where the combined value is unknown	Item 14	Medium
Able to numerically map a small 'change: start unknown' subtraction word problem	Mapping a subtraction word problem with two numbers where the starting value is unknown	Item 20	Medium
Able to numerically map a complex 'combine: combine value unknown' addition word problem	Mapping an addition word problem with three numbers where numbers have to be combined in two steps	Item 26	High
Able to numerically map a complex 'compare: compared quality unknown' subtraction word problem	Mapping a subtraction word problem with three numbers where two steps are necessary to find the compared quality	Item 32	High

Selecting an appropriate strategy for solution:

Construct map and items for strand 3: Selecting an appropriate strategy for solution			
Ability level	Content specification	Test Item Number	Item Difficulty level
Able to add two large numbers with carrying	Addition of two four-digit numbers using carrying once	Item 3	Low
Able to subtract two large numbers without borrowing	Subtracting two four-digit numbers without borrowing	Item 9	Low
Able to add two large numbers where carrying is required twice	Addition of two five-digit numbers using carrying twice	Item 15	Medium
Able to subtract two large numbers with borrowing	Subtracting two four-digit numbers using borrowing once	Item 21	Medium
Able to add three large numbers with carrying three times	Addition of three five-digit numbers using carrying three times	Item 27	High
Able to do subtraction involving three numbers where addition is required	Subtracting three four-digit numbers where two smaller numbers need to be added first	Item 33	High

Performing the calculations on the stored numbers:

Construct map and items for strand 4: Performing the calculations on the stored numbers			
Ability level	Content specification	Test Item Number	Item Difficulty level
Able to add two small numbers	Addition of two numbers (maximum three digits)	Item 4	Low
Able to subtract two small numbers	Subtracting two numbers (maximum three-digits)	Item 10	Low
Able to add two small numbers approximating for the result	Addition of two numbers (maximum two digits) and approximation for result	Item 16	Medium
Able to subtract two small numbers approximating for the result	Subtracting two numbers (maximum two digits) and approximation for result	Item 22	Medium
Able to add two large numbers approximating for result	Addition of two numbers and approximation for result, maximum four digits	Item 28	High
Able to subtract two large numbers	Subtracting two numbers, both four digits	Item 34	High

Verifying the solution and communicating the results:

Construct map and items for strand 5: Verifying the solution and communicating the results			
Ability level	Content specification	Test Item Number	Item Difficulty level
Able to verify solution and result of a small addition	Verifying the solution and communicating the result for adding two numbers (maximum three-digit number)	Item 5	Low
Able to verify solution and result of a small subtraction	Verifying the solution and communicating the result for subtracting two numbers (maximum three-digit number)	Item 11	Low
Able to verify solution and result for a large addition	Verifying the solution and communicating the result for adding three numbers (maximum four-digit number)	Item 17	Medium
Able to verify solution and result for a large subtraction	Verifying the solution and communicating the result for subtracting two numbers (maximum four-digit number)	Item 23	Medium
Able to verify solution and result for a large addition while applying real-life understanding	Verifying the solution and communicating the result for an addition with understanding of real-life situation	Item 29	High
Able to verify solution and result for a small subtraction while applying real-life understanding	Verifying the solution and communicating the result for a subtraction with understanding of real-life situation	Item 35	High
Able to verify solution and result with knowledge of time-format expression	Verifying the solution and the result based on correct format for expressing time	Item 37	High
Able to verify solution and result with knowledge of unit of currency	Verifying the solution and the result based on correct format for unit of currency	Item 38	High

Appendix 3.3: Pilot study test booklet (English)

Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

Student school S/N:
School research ID:
Student research ID:

Grade (tick as appropriate): (a) Class 4 (b) Class 5

You are a? (tick as appropriate): (a) Girl (b) Boy

Year of birth (please write): _____

Have you attended pre-primary (tick as appropriate): (a) Yes (b) No

Do you attend private tuition after school? (Tick as appropriate): (a) Yes (b) No

Parents' education (if known to you) Mother: _____ Father: _____

Parents' profession (if known to you) Mother: _____ Father: _____

Family's average monthly expenditure (if known to you): _____

(Answer the following items by choosing the correct alternative by putting a tick mark (✓) on it, if you want you can use the attached additional paper for calculations. One example has been done for you)

Example: Akash has 5 toys. Ria has 3 toys. How many toys do they have together?

- a. 5
- b. 3
- ✓ c. 8
- d. 2
- e. 4
- f. none of the above

1. Jahid had 5 apples, and then Koli gave him 7 apples. How many apples did Jahid have in the beginning?
a. 12 b. 7 c. 5 d. 2 e. Not enough information to tell
2. Razia has 15 candles. Daud has 7 candles less than Razia. Which of these number sentences would be appropriate for calculating how many candles Daud has?
a. $15 + 7$
b. $7 + 15$
c. $15 - 7$
d. $7 - 15$
e. None of the above

3. Solve this addition problem $24 + 31$. Show your work.
4. $22 + 7 = ?$
- 27
 - 24
 - 29
 - 30
 - None of the above
5. Suppose you were asked to add 235 and 54, and found the result was 289. How would you write the solution down?
- $$\begin{array}{r} 235 \\ 54+ \\ \hline 289 \end{array}$$
 - $$\begin{array}{r} 289 \\ +54 \\ \hline 235 \end{array}$$
 - $$\begin{array}{r} 54+ \\ 234 \\ \hline 289 \end{array}$$
 - $$\begin{array}{r} 235 \\ +54 \\ \hline 289 \end{array}$$
 - None of the above
6. In a cricket match Sujon got 120 runs, Sumon got 85 runs, and Sujoy got 67 runs. How many runs did they score together?
- 205
 - 272
 - 152
 - 372
 - None of the above
7. Three of Raju's friends gave him some coins. First, Samia gave him 45 coins; then Jakaria gave him 80 coins; and finally Lisa gave him 25 coins. How many coins did he get from the last one of his friends?
- 125
 - 80
 - 25
 - 45
 - None of the above

8. Latif has 138 plants in his house. Mitu has 112 plants more than him. Which of these number sentences would be appropriate for calculating how many plants Mitu has?

- a. $138 - 12$
- b. $112 - 138$
- c. $140 - 12$
- d. $140 - 138$
- e. None of the above

9. Solve this subtraction problem $26 - 12$. Show your work.

10. $23 - 8 = ?$

- a. 15
- b. 31
- c. 17
- d. 16
- e. None of the above

11. Suppose you were asked to subtract 84 from 439 and you found the answer was 355. How would you write the solution down?

a.
$$\begin{array}{r} 84 \\ - 439 \\ \hline 355 \end{array}$$

b.
$$\begin{array}{r} 439 \\ - 84 \\ \hline 355 \end{array}$$

c.
$$\begin{array}{r} 439 \\ 84 - \\ \hline 355 \end{array}$$

d.
$$\begin{array}{r} 355 \\ + 84 \\ \hline 439 \end{array}$$

- e. None of the above

12. Mr. Ajay went to the market with 4250 taka. After doing shopping he had 890 taka left with him. How much money did he spend in the market?
- 5140
 - 890
 - 3360
 - 4210
 - None of the above
13. Jamal has 5 toys, Bakar has 8 toys. Which of the following is true?
- Jamal has more toys than Bakar
 - Bakar has more toys than Jamal
 - Bakar has less toys than Jamal
 - Jamal and Bakar have equal numbers of toys
 - None of the above
14. Two hundred and twenty-three boys, one hundred and thirty-five girls, and twelve teachers participated in a school trip. Which of these number sentences is correct to find out the total number of people going on the trip?
- $223 + 135$
 - $135 + 12 - 223$
 - $12 + 135 + 223$
 - $223 - 135 - 12$
 - None of the above
15. Solve this addition problem $58 + 27$. Show your work.
16. Which of the following is the closest to the result for solving $94 + 33$?
- 110
 - 130
 - 120
 - 140
 - None of the above
17. Suppose you were asked to add 26, 2533, and 325. You found the answer was 2884. How would you write the solution down?
- $$\begin{array}{r}
 2433 \\
 26 \\
 +325 \\
 \hline
 2884
 \end{array}$$

$$\begin{array}{r} \text{b. } 2533 \\ 325 \\ - 26 \\ \hline 2884 \end{array}$$

$$\begin{array}{r} \text{c. } 2533 \\ 325 \\ + 26 \\ \hline 2884 \end{array}$$

$$\begin{array}{r} \text{d. } 2533 \\ 325+ \\ 26 \\ \hline 2884 \end{array}$$

e. None of the above

18. In a city, there live 23465 male, 2335 female and 11289 children. What is the number of population of the city?

- a. 37080
- b. 40830
- c. 35780
- d. 36900
- e. None of the above

19. A cricket bat weighs 2 kilograms, a pencil weighs 50 grams, a ruler weighs 100 grams, a television weighs 6 kilograms, and a tomato weighs 30 grams. A cricket bat has the highest length among these but which of these objects weighs more than the cricket bat?

- a. Pencil
- b. Ruler
- c. Tomato
- d. Television
- e. None of the above

20. Joba has some oranges. Dilu has 28 oranges which is 12 more than Joba. Which of the following number sentences is right for figuring out how many oranges Joba has?

- a. $28 + 12$
- b. $12 + 28$
- c. $28 - 12$
- d. $40 - 12$
- e. None of the above

21. Solve this subtraction problem $63 - 19$. Show your work.

22. Which of the following is the closest to the result for solving $87 - 53$?

- a. 37
- b. 45
- c. 27
- d. 30
- e. None of the above

23. Suppose you were asked to subtract 129 from 2613 and you found the answer was 2484. How would you write the solution down?

a.
$$\begin{array}{r} 2631 \\ -129 \\ \hline 2484 \end{array}$$

b.
$$\begin{array}{r} -129 \\ \hline 2613 \\ 2884 \end{array}$$

c.
$$\begin{array}{r} 2613 \\ -129 \\ \hline 2484 \end{array}$$

d.
$$\begin{array}{r} 2613 \\ 129- \\ \hline 2484 \end{array}$$

- e. None of the above

24. A school needs 9500 taka for annual sports. The school got 4500 taka as government grant and 2000 taka from school fund. How much more money do they need to collect?

- a. 4500
- b. 2000
- c. 3000
- d. 5000
- e. None of the above

25. Khulna is to the west from Dhaka, Narsingdi is to the east of Dhaka. Sylhet is further east from Dhaka. Which of these cities is furthest towards the west?
- a. Sylhet
 - b. Dhaka
 - c. Narsingdi
 - d. Khulna
 - e. Not enough information to tell
26. Rahat has 41 stamps, Rubin has 30 stamps. Lisa has 23 stamps more than what Rahat and Rubin has together. Which of the following number sentences is right for calculating how many stamps Lisa has?
- a. $41 + 30 - 23$
 - b. $41 + 30 + 23$
 - c. $30 - 23 + 41$
 - d. $41 - 30 - 23$
 - e. None of the above
27. Solve this addition $1039 + 142 + 14$. Show your work.
28. Which of the following is the closest to the result for solving for $2137 + 251$?
- a. 2350
 - b. 2360
 - c. 2400
 - d. 2420
 - e. None of the above
29. A teacher is buying some new furniture for his school. He bought 7 desks. 2 of these are 7 feet long and weighs 25 kilo each. The other 5 are 5 feet long and weighs 20 kilo each. The teacher calculated the total weight of the desks to be 150 kg and arranged for a transportation. Which of the following alternatives show the accurate calculation and the appropriate transportation?
- a. Weight $25+25+25+25+20+20+20= 150$ kg; a rickshaw for transportation
 - b. Weight $25+25+25+25+20+20+20= 150$ kg; a rickshaw van for transportation
 - c. Weight $25+25+20+20+20+20+20 = 150$ kg; a truck for transportation

- d. Weight $25+25+20+20+20+20+20 = 150$ kg; a rickshaw van for transportation
 - e. Weight $25+25+20+20+20+20+20 = 150$ kg; a bus for transportation
30. The number of males and females in a town is 42530. Among them, the number of males is 21678; find the number of females.
- a. 19561
 - b. 20852
 - c. 20521
 - d. 19322
 - e. None of the above
31. Water from a swimming pool has some salt in it. Tap water has the least salt. But there is much salt in water from the Atlantic or the Pacific Ocean. Which water contains the most salt?
- a. Tap water
 - b. Swimming pool water
 - c. Water from Atlantic
 - d. Water from the Pacific
 - e. Not enough information to tell
32. Rubi got 2500 taka from her parents to buy new books. Nusrat got 700 taka less than Rubi and Rima got 300 taka less than Nusrat. Which of the following number sentences is correct for calculating the amount Rima got?
- a. $2500 + 700 + 300$
 - b. $2500 - 700 + 300$
 - c. $2500 - 700 - 300$
 - d. $700 + 300 - 2500$
 - e. None of the above
33. Solve this subtraction problem $1269 - 233 - 12$. Show your work.
34. Which of the following is the correct answer for solving $1245 - 118$?
- a. 1127
 - b. 1117
 - c. 1137

- d. 1023
 - e. None of the above
35. The headteacher of a school is planning a victory day celebration in her school and wanted each student to have a small Bangladeshi flag. These flags come in packets of 200s. The school has 656 students in total but 126 of them will not be present. The headteacher calculated there will be 530 students in the celebration and the school has 4 packets of flag. Which of the following alternatives show the correct calculation and result?
- a. $656 + 126 = 530$; 4 packets of flag
 - b. $656 - 126 = 530$; 4 packets of flag
 - c. $656 - 126 = 530$; 3 packets of flag
 - d. $656 - 126 = 530$; 2 packets of flag
 - e. $656 + 126 = 530$; 3 packets of flag
36. Shihab has 390 taka more than that of Siam and Siam has 470 taka less than that of Shimul. Shimul has 890 taka. How much money does each of Siam and Shihab has?
- a. Siam 470 taka; Shihab 390 taka
 - b. Siam 810 taka; Shihab 420 taka
 - c. Siam 420 taka; Shihab 810 taka
 - d. Siam 860 taka; Shihab 890 taka
 - e. None of the above
37. Pia was asked to solve this problem:
 'Rashed started for his school from home at 10.40 am and reached the class at 11.30 am. How long did it take for him to reach his class?'
 She solved the problem and found the answer was 50. How shall she write down the result?
- a. 50 am
 - b. 50 hours
 - c. 50 minutes
 - d. 0.50 hours
 - e. None of the above
38. Leon was asked to solve the following problem:
 'Jenny bought 5 candies for 12.5 taka. She gave the shopkeeper a 20-taka note. How much shall she get in change?'
 He solved the problem and found the answer was 7.5. How should he write down the answer?
- a. 7 taka
 - b. 7 taka and 50 poisa
 - c. 7.5 poisa

- d. 7.5
- e. None of the above

39. Think about how you would solve for this addition: $25+32$. Which of the following most closely matches your thinking process?

- a. you do column addition

$$\begin{array}{r} 25 \\ +32 \\ \hline 57 \end{array}$$

- b. you remember $25 + 32 = 57$
- c. first you do $20 + 30 = 50$, then $5+2= 7$, and then $50+7=57$
- d. first you do $32 + 8 = 40$, then $25 - 8 = 17$; and then $40 + 17 = 57$
- e. you are not sure how to solve this problem
- f. you are going to solve it in some other way. Please write down here:

40. Think about how you would solve for this problem: $14 - 8$. Which of the following most closely matches your thinking process?

- a. you guess/you don't know how to solve this problem
- b. you remember $14 - 8 = 6$
- c. you count up from 8 till you get to 14 (8, 9, 10,...,14) to get the answer 6
- d. you remember $8 = 4 + 4$, and $14 - 4 = 10$, and then $10 - 4 = 6$
- e. you will do columnar subtraction

$$\begin{array}{r} 14 \\ -8 \\ \hline 6 \end{array}$$

- f. some other method – please write an explanation of your method:

(Thanks for participating in this test)

Appendix 3.4: Pilot study test booklet (Bangla)

Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

শিক্ষার্থীর বিদ্যালয় রোলঃ
বিদ্যালয়ের গবেষণা আইডিঃ
শিক্ষার্থীর গবেষণা আইডিঃ

তুমি কোন শ্রেণীতে পড়? (টিক চিহ্ন দাও) : (ক) ৪র্থ শ্রেণী (খ) ৫ম শ্রেণী

তুমি কী? (টিক চিহ্ন দাও) (ক) ছাত্রী (খ) ছাত্র

তোমার জন্মসাল (পাশের খালি জায়গায় লিখ)

তুমি কি শিশু শ্রেণীতে পড়েছ? (টিক চিহ্ন দাও) (ক) হ্যাঁ (খ) না

তুমি কি স্কুল সময়ের পরে কোন শিক্ষকের কাছে প্রাইভেট পড়? (টিক চিহ্ন দাও) (ক) হ্যাঁ (খ) না

তোমার অভিভাবকদের শিক্ষাগত যোগ্যতা (জানা থাকলে লিখ) : মা বাবা

তোমার অভিভাবকদের পেশা (জানা থাকলে লিখ) : মা বাবা

তোমার পরিবারের গড় মাসিক ব্যয় (জানা থাকলে লিখ) : (টাকা)

(নিচের প্রশ্নগুলোতে যেখানে প্রযোজ্য সঠিক উত্তরে টিক চিহ্ন দাও। কিছু প্রশ্নের নিচে দেয়া খালি জায়গায় উত্তর লিখতে হবে। তুমি চাইলে হিসেব করার জন্য অলান্দা কাগজ ব্যবহার করতে পার। এখানে একটি উদাহরণের সমাধান করা আছে।)

উদাহরণঃ

আকাশের ৫ টি খেলনা আছে। রিয়ার ৩ টি খেলনা আছে। তাদের একত্রে কয়টি খেলনা আছে?

ক) ৫

খ) ৩

✓ গ) ৮

ঘ) ২

ঙ) ৪

০১) জাহিদের কাছে ৫ টা আপেল ছিল। কলি তাকে আরো ৭ টা আপেল দিল। জাহিদের কাছে শুরুতে ক'টা আপেল ছিল?

(ক) ১২ (খ) ৭ (গ) ৫ (ঘ) ২ (ঙ) উত্তর দেবার মত যথেষ্ট তথ্য নেই

০২) রাজিয়ার কাছে ১৫ টি মোমবাতি আছে। দাউদের কাছে রাজিয়ার চেয়ে ৭ টি মোমবাতি কম আছে। দাউদের কাছে কতগুলো মোমবাতি আছে সেটা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

(ক) $১৫ + ৭$

(খ) $৭ + ১৫$

(গ) $১৫ - ৭$

(ঘ) $৭ - ১৫$

(ঙ) উপরের কোনটিই নয়

০৩) নিচের খালি জায়গায় এই যোগ অঙ্কটি করঃ $২৪ + ৩১$ ।

০৪) $২২ + ৭ = ?$

- (ক) ২৭
- (খ) ২৪
- (গ) ২৯
- (ঘ) ৩০
- (ঙ) উপরের কোনটিই নয়

০৫) মনে কর তোমাকে বলা হল ২৩৫ আর ৫৪ যোগ করতে। যোগ করে তুমি উত্তর পেলে ২৮৯ । এই যোগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

(ক) ২৩৫
 $\begin{array}{r} ৫৪+ \\ ২৮৯ \end{array}$

(খ) ২৮৯
 $\begin{array}{r} +৫৪ \\ ২৩৫ \end{array}$

(গ) $৫৪+$
 $\begin{array}{r} ২৩৫ \\ ২৮৯ \end{array}$

(ঘ) ২৩৫
 $\begin{array}{r} +৫৪ \\ ২৮৯ \end{array}$

- (ঙ) উপরের কোনটিই নয়

০৬) ক্রিকেট খেলায় সুজন ১২০ রান, সুমন ৮৫ রান এবং সুজয় ৬৭ রান করল। তিনজনে মোট কত রান করল?

- (ক) ২০৫
- (খ) ২৭২
- (গ) ১৫২
- (ঘ) ৩৭২
- (ঙ) উপরের কোনটিই নয়

০৭) রাজুর তিনজন বন্ধু তাকে কিছু পয়সা দিল। প্রথমে সামিয়া তাকে ৪৫ টি পয়সা দিল, জাকারিয়া তাকে ৮০ টি পয়সা দিল এবং তারপর লিসা তাকে ২৫ টি পয়সা দিল। বন্ধুদের মধ্যে শেষজন রাজুকে কয়টি পয়সা দিয়েছিল?

- (ক) ১২৫
- (খ) ৮০
- (গ) ২৫
- (ঘ) ৪৫
- (ঙ) কোনটিই নয়

০৮) লতিফের বাসায় ১৩৮ টি ফুলগাছ আছে। মিতুর বাসায় লতিফের চেয়ে ১১২ টি ফুলগাছ বেশী আছে। মিতুর বাসায় কতগুলো ফুলগাছ আছে সেটা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $১৩৮ - ১২$
- (খ) $১১২ - ১৩৮$
- (গ) $১৪০ - ১২$
- (ঘ) $১৪০ - ১৩৮$
- (ঙ) উপরের কোনটিই নয়

০৯) নিচের খালি জায়গায় এই বিয়োগ অঙ্কটি করঃ $২৬ - ১২$ ।

১০) $২৩ - ৮ = ?$

- (ক) ১৫
- (খ) ৩১
- (গ) ১৭
- (ঘ) ১৬
- (ঙ) উপরের কোনটিই নয়

১১) মনে কর, তোমাকে বলা হল ৪৩৯ থেকে ৮৪ বিয়োগ করতে। বিয়োগ করে উত্তর পাওয়া গেল ৩৫৫। এই বিয়োগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

(ক) ৮৪

$$\begin{array}{r} -84 \\ 439 \\ \hline 355 \end{array}$$

(খ) ৪৩৯

$$\begin{array}{r} -84 \\ 439 \\ \hline 355 \end{array}$$

(গ) ৪৩৯

$$\begin{array}{r} 84- \\ 439 \\ \hline 355 \end{array}$$

(ঘ) ৩৫৫

$$\begin{array}{r} +84 \\ 355 \\ \hline 439 \end{array}$$

(ঙ) উপরের কোনটিই নয়

১২) অজয় বাবু ৪২৫০ টাকা নিয়ে বাজারে গেলেন। বাজার করার পর তার কাছে ৮৯০ টাকা রইল। তিনি কত টাকার বাজার করলেন?

(ক) ৫১৪০

(খ) ৮৯০

(গ) ৩৩৬০

(ঘ) ৪২১০

(ঙ) উপরের কোনটিই নয়

১৩) জামালের কাছে ৫ টি খেলনা আছে, বকরের কাছে ৮ টি খেলনা আছে। নিচের কোনটি সত্য?

(ক) জামালের বকরের চেয়ে বেশী খেলনা আছে

(খ) বকরের জামালের চেয়ে বেশী খেলনা আছে

(গ) বকরের জামালের চেয়ে কম খেলনা আছে

(ঘ) জামাল আর বকরের সমান সংখ্যক খেলনা আছে

(ঙ) উপরের কোনটিই নয়

১৪) একটি স্কুল থেকে দুইশ তেইশজন ছাত্র, একশ পঁয়ত্রিশ জন ছাত্রী এবং বারজন শিক্ষক শিক্ষাসফরে যাচ্ছেন। মোট কতজন শিক্ষাসফরে যাচ্ছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

(ক) ২২৩ + ১৩৫

(খ) ১৩৫ + ১২ - ২২৩

(গ) ১২ + ১৩৫ + ২২৩

- (ঘ) ২২৩ - ১৩৫ - ১২
(ঙ) উপরের কোনটিই নয়

১৫) নিচের খালি জায়গায় এই যোগ অঙ্কটি করঃ ৫৮ + ২৭।

১৬) নিচের কোন বিকল্পটি ৯৪ + ৩৩ এর যোগফলের সবচেয়ে কাছাকাছি?

- (ক) ১১০
(খ) ১৩০
(গ) ১২০
(ঘ) ১৪০
(ঙ) উপরের কোনটিই নয়

১৭) মনে কর তোমাকে বলা হল ২৬, ২৫৩৩ আর ৩২৫ যোগ করতে। যোগ করে তুমি উত্তর পেলে ২৮৮৪। এই যোগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

(ক) ২৪৩৩
২৬
+ ৩২৫

২৮৮৪

(খ) ২৫৩৩
৩২৫
- ২৬

২৮৮৪

(গ) ২৫৩৩
৩২৫
+ ২৬

২৮৮৪

(ঘ) ২৫৩৩
৩২৫+
২৬

২৮৮৪

- (ঙ) উপরের কোনটিই নয়

১৮) একটি শহরে ২৩৪৬৫ জন পুরুষ, ২৩৩৫ জন মহিলা ও ১১২৮৯ জন শিশু বাস করে। উক্ত শহরের জনসংখ্যা কত?

- (ক) ৩৭০৮০
- (খ) ৪০৮৩০
- (গ) ৩৫৭৮০
- (ঘ) ৩৬৯০০
- (ঙ) উপরের কোনটিই নয়

১৯) একটি ক্রিকেট ব্যাটের ওজন ২ কিলোগ্রাম, একটি পেসিলের ওজন ৫০ গ্রাম, একটি ব্ললারের ওজন ১০০ গ্রাম, একটি টেলিভিশনের ওজন ৬ কিলোগ্রাম, এবং একটি টমেটোর ওজন ৩০ গ্রাম। এগুলোর মধ্যে ক্রিকেট ব্যাটটি লম্বায় সবচেয়ে বড় কিন্তু কোন বস্তুটির ওজন ক্রিকেট ব্যাটের চেয়ে বেশী?

- (ক) পেসিলা
- (খ) ব্ললার
- (গ) টমেটো
- (ঘ) টেলিভিশন
- (ঙ) উপরের কোনটিই নয়

২০) জবার কাছে কিছু কমলালেবু আছে। দিলুর কাছে ২৮ টি কমলালেবু আছে যা জবার চেয়ে ১২ টি বেশী। জবার কাছে ক'টি কমলালেবু আছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $২৮ + ১২$
- (খ) $১২ + ২৮$
- (গ) $২৮ - ১২$
- (ঘ) $৪০ - ১২$
- (ঙ) উপরের কোনটিই নয়

২১) নিচের খালি জায়গায় এই বিয়োগ অঙ্কটি করঃ ৬৩ - ১৯ ।

২২) নিচের কোন বিকল্পটি ৮৭ - ৫৩ এর বিয়োগফলের সবচেয়ে কাছাকাছি?

- (ক) ৩৭
- (খ) ৪৫
- (গ) ২৭
- (ঘ) ৩০
- (ঙ) উপরের কোনটিই নয়

২৩) মনে কর, তোমাকে বলা হল ২৬১৩ থেকে ১২৯ বিয়োগ করতে। বিয়োগ করে উত্তর পাওয়া গেল ২৪৮৪। এই বিয়োগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

(ক) ২৬৩১

$$\begin{array}{r} -129 \\ 2613 \\ \hline 2484 \end{array}$$

(খ) - ১২৯

$$\begin{array}{r} 2613 \\ -129 \\ \hline 2484 \end{array}$$

(গ) ২৬১৩

$$\begin{array}{r} -129 \\ 2613 \\ \hline 2484 \end{array}$$

(ঘ) ২৬১৩

$$\begin{array}{r} 129- \\ 2613 \\ \hline 2484 \end{array}$$

(ঙ) উপরের কোনটিই নয়

২৪) বার্ষিক ক্রীড়া প্রতিযোগিতার জন্য ৯৫০০ টাকা প্রয়োজন। সরকারি অনুদান হিসেবে ৪৫০০ টাকা এবং বিদ্যালয়ের তহবিল থেকে ২০০০ টাকা দেয়া হয়েছে। অনুষ্ঠানটি করার জন্য আর কত টাকা যোগাড় করতে হবে?

(ক) ৪৫০০

(খ) ২০০০

(গ) ৩০০০

(ঘ) ৫০০০

(ঙ) উপরের কোনটিই নয়

২৫) খুলনা ঢাকা থেকে পশ্চিম দিকে অবস্থিত, নরসিংদী ঢাকা থেকে পূর্ব দিকে অবস্থিত। সিলেট ঢাকা থেকে আরও পূর্ব দিকে অবস্থিত। এই শহরগুলোর মধ্যে কোনটি সর্বাধিক পশ্চিমে অবস্থিত?

(ক) সিলেট

(খ) ঢাকা

(গ) নরসিংদী

(ঘ) খুলনা

(ঙ) উত্তর দেবার মত যথেষ্ট তথ্য নেই

২৬) রাহাতের কাছে ৪১ টি ডাকটিকিট আছে, রুবিনের কাছে ৩০ টি ডাকটিকিট আছে। লিসার কাছে রাহাত আর রুবিন দু'জনের মিলে যতগুলো ডাকটিকিট তার চেয়ে ২৩ টি ডাকটিকিট বেশী আছে। লিসার কাছে কতটি ডাকটিকিট আছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $৪১ + ৩০ - ২৩$
- (খ) $৪১ + ৩০ + ২৩$
- (গ) $৩০ - ২৩ + ৪১$
- (ঘ) $৪১ - ৩০ - ২৩$
- (ঙ) উপরের কোনটিই নয়

২৭) নিচের খালি জায়গায় এই যোগ অঙ্কটি করঃ $১০৩৯ + ১৪২ + ১৪$ ।

২৮) নিচের কোন বিকল্পটি $২১৩৭ + ২৫১$ এর যোগফলের সবচেয়ে কাছাকাছি?

- (ক) ২৩৫০
- (খ) ২৩৬০
- (গ) ২৪০০
- (ঘ) ২৪২০
- (ঙ) উপরের কোনটিই নয়

২৯) একজন শিক্ষক তার স্কুলের জন্য নতুন আসবাবপত্র কিনেছেন। তিনি দু'টো টেবিল কিনেছেন যেগুলোর প্রতিটির দৈর্ঘ্য ৭ ফুট এবং ওজন ২৫ কিলোগ্রাম। এছাড়াও তিনি ৫ টি বেঞ্চ কিনেছেন যেগুলোর প্রতিটির দৈর্ঘ্য ৫ ফুট এবং ওজন ২০ কিলোগ্রাম। শিক্ষক হিসেব করে দেখলেন আসবাবপত্রগুলোর মোট ওজন ১৫০ কিলোগ্রাম এবং সেগুলোকে স্কুলে নিয়ে যাবার জন্য একটি বাহনের ব্যবস্থা করলেন। নিচের বিকল্পগুলোর মধ্যে কোনটিতে সবচেয়ে সঠিক হিসেব এবং উপযুক্ত পরিবহন মাধ্যম রয়েছে?

- (ক) ওজন $২৫ + ২৫ + ২৫ + ২৫ + ২০ + ২০ + ২০ = ১৫০$ কিলোগ্রাম; পরিবহনের জন্য রিকশা
- (খ) ওজন $২৫ + ২৫ + ২৫ + ২৫ + ২০ + ২০ + ২০ + ২০ = ১৫০$ কিলোগ্রাম; পরিবহনের জন্য রিকশা ভ্যান
- (গ) ওজন $২৫ + ২৫ + ২০ + ২০ + ২০ + ২০ + ২০ = ১৫০$ কিলোগ্রাম; পরিবহনের জন্য ট্রাক
- (ঘ) ওজন $২৫ + ২৫ + ২০ + ২০ + ২০ + ২০ + ২০ = ১৫০$ কিলোগ্রাম; পরিবহনের জন্য বাস
- (ঙ) উপরের কোনটিই নয়

৩০) কোন শহরে নারী পুরুষের সংখ্যা ৪২৫৩০, এদের মধ্যে ২১৬৭৮ জন পুরুষ হলে, নারীর সংখ্যা নির্ণয় কর।

- (ক) ১৯৫৬১
- (খ) ২০৮৫২
- (গ) ২০৫২১
- (ঘ) ১৯৩২২
- (ঙ) উপরের কোনটিই নয়

৩১) সুইমিং পুলের পানিতে খানিকটা লবণ থাকে। কলের পানিতে সবচেয়ে কম লবণ থাকে। তবে আটলান্টিক মহাসাগর বা প্রশান্ত মহাসাগরের পানিতে প্রচুর লবণ থাকে। কোন পানিতে সবচেয়ে বেশী লবণ থাকে?

- (ক) কলের পানিতে
- (খ) সুইমিং পুলের পানিতে
- (গ) আটলান্টিক মহাসাগরের পানিতে
- (ঘ) প্রশান্ত মহাসাগরের পানিতে
- (ঙ) উত্তর দেবার মত যথেষ্ট তথ্য নেই

৩২) নতুন ক্লাসের বই-খাতা কেনার জন্য রুবি বাসা থেকে ২৫০০ টাকা পেয়েছে। নুসরাত রুবির চেয়ে ৭০০ টাকা কম পেয়েছে আর রিমা নুসরাতের চেয়ে ৩০০ টাকা কম পেয়েছে। রিমা কত টাকা পেয়েছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $২৫০০ + ৭০০ + ৩০০$
- (খ) $২৫০০ - ৭০০ + ৩০০$
- (গ) $২৫০০ - ৭০০ - ৩০০$
- (ঘ) $৭০০ + ৩০০ - ২৫০০$
- (ঙ) উপরের কোনটিই নয়

৩৩) নিচের খালি জায়গায় এই বিয়োগ অঙ্কটি করঃ ১২৬৯ - ২৩৩ - ১২।

৩৪) ১২৪৫ - ১১৮ এর বিয়োগফলের জন্য সঠিক উত্তর নিচের কোনটি?

- (ক) ১১২৭
- (খ) ১১১৭
- (গ) ১১৩৭
- (ঘ) ১০২৩
- (ঙ) উপরের কোনটিই নয়

৩৫) একটি স্কুলের প্রধান শিক্ষক স্কুলে বিজয় দিবস উদযাপনের আয়োজন করতে চাইছেন যেখানে সব ছাত্র-ছাত্রীর হাতে বাংলাদেশের একটি করে ছোট পতাকা থাকবে। এই পতাকাগুলো ২০০ টির প্যাকেটে পাওয়া যায়। স্কুলে এমন ৪ টি প্যাকেট আছে। ওই স্কুলে মোট ৬৫৬ জন ছাত্র-ছাত্রী আছে কিন্তু তাদের মধ্যে ১২৬ জন বিজয় দিবসের অনুষ্ঠানে উপস্থিত থাকতে পারবে না। প্রধান শিক্ষক হিসেব করে দেখলেন অনুষ্ঠানে মোট ৫৩০ জন ছাত্র-ছাত্রী থাকবে এবং নতুন করে পতাকাও কিনতে হবে না। নিচের কোন বিকল্পটিতে অনুষ্ঠানে উপস্থিত ছাত্র-ছাত্রী সংখ্যা ও প্রয়োজনীয় পতাকার সংখ্যা বের করার জন্য সঠিক গণনা এবং উত্তর আছে?

- (ক) $৬৫৬ + ১২৬ = ৫৩০$; ৪ প্যাকেট পতাকা
- (খ) $৬৫৬ - ১২৬ = ৫৩০$; ৪ প্যাকেট পতাকা
- (গ) $৬৫৬ - ১২৬ = ৫৩০$; ৩ প্যাকেট পতাকা
- (ঘ) $৬৫৬ - ১২৬ = ৫৩০$; ২ প্যাকেট পতাকা
- (ঙ) উপরের কোনটিই নয়

৩৬) সিয়াম অপেক্ষা শিহাবের ৩৯০ টাকা বেশি এবং শিমুল অপেক্ষা ৪৭০ টাকা কম আছে। শিমুলের ৮৯০ টাকা আছে। সিয়াম ও শিহাবের প্রত্যেকের কত টাকা আছে?

- (ক) সিয়াম ৪৭০ টাকা, শিহাব ৩৯০ টাকা
- (খ) সিয়াম ৮১০ টাকা, শিহাব ৪২০ টাকা
- (গ) সিয়াম ৪২০ টাকা, শিহাব ৮১০ টাকা
- (ঘ) সিয়াম ৮৬০ টাকা, শিহাব ৮৯০ টাকা
- (ঙ) উপরের কোনটিই নয়

৩৭) পিয়াকে নিচের অঙ্কটি করতে দেয়া হয়েছিল:

‘রাশেদ স্কুলে যাবার জন্য বাসা থেকে সকাল ১০.২০ এ বের হয়ে সকাল ১১.৩০ এ ক্লাসে পৌঁছাল। তার ক্লাসে পৌঁছতে কত সময় লেগেছে?’

পিয়া অঙ্কটি করে উত্তর পেল ৫০। উত্তরটা সে কিভাবে লিখবে?

- (ক) সকাল ৫০
- (খ) ৫০ ঘন্টা
- (গ) ৫০ মিনিট
- (ঘ) ০.৫০ ঘন্টা
- (ঙ) উপরের কোনটিই নয়

৩৮) লিয়নকে নিচের অঙ্কটি করতে দেয়া হয়েছিল:

'জেনি ১২.৫ টাকা দিয়ে ৫ টি চকলেট কিনেছে। সে দোকানীকে ২০ টাকার একটা নোট দিল। দোকানী তাকে কত ফেরত দেবে?'

লিয়ন অঙ্কটা করে উত্তর পেল ৭.৫। উত্তরটা সে কিভাবে লিখবে?

- (ক) ৭ টাকা
- (খ) ৭ টাকা ৫০ পয়সা
- (গ) ৭.৫ পয়সা
- (ঘ) ৭.৫
- (ঙ) উপরের কোনটিই নয়

৩৯) $২৫ + ৩২$ এই যোগ অঙ্কটি তুমি কিভাবে করবে চিন্তা কর। নিচের কোন বিকল্পটি তোমার চিন্তার সাথে সবচেয়ে বেশী মেলে?

(ক) তুমি উপর-নিচে যোগ করবে

$$\begin{array}{r} ২৫ \\ + ৩২ \\ \hline ৫৭ \end{array}$$

- (খ) $২৫ + ৩২ = ৫৭$ হয় এটা তুমি মুখস্থ জানো
- (গ) তুমি মনে মনে হিসেব করবে যে $২০ + ৩০ = ৫০$ আর $৫ + ২ = ৭$; এর পরে $৫০ + ৭ = ৫৭$
- (ঘ) তুমি মনে মনে হিসেব করবে যে $৩২ + ৮ = ৪০$ আর $২৫ - ৮ = ১৭$; এর পরে $৪০ + ১৭ = ৫৭$
- (ঙ) এই অঙ্কটি কিভাবে করতে হবে তুমি নিশ্চিত নও
- (চ) তুমি অন্য কোন পদ্ধতি ব্যবহার করবে। কী পদ্ধতি তা নিচের খালি জায়গায় লিখ:

৪০) $18 - c$ এই বিয়োগ অঙ্কটি তুমি কিভাবে করবে চিন্তা কর। নিচের কোন বিকল্পটি তোমার চিন্তার সাথে সবচেয়ে বেশী মেলে?

- (ক) তুমি আন্দাজে উত্তর দেবে/ অঙ্কটি করার পদ্ধতি তোমার জানা নেই
(খ) $18 - c = 6$ হয় এটা তুমি মুখস্থ জানো
(গ) তুমি c থেকে $c, 9, 10$ এভাবে 18 পর্যন্ত গুণতে থাকবে এবং উত্তর 6 হয় সেটা বের করবে
(ঘ) তুমি জানো $c = 8 + 8$ আর $18 - 8 = 10$, এরপর $10 - 8 = 6$
(ঙ) তুমি উপর-নিচে বিয়োগ করবে

$$\begin{array}{r} 18 \\ - c \\ \hline 6 \end{array}$$

- (চ) তুমি অন্য কোন পদ্ধতি ব্যবহার করবে। কী পদ্ধতি তা নিচের খালি জায়গায় লিখ:

(সহযোগিতার জন্য তোমাকে ধন্যবাদ)

Appendix 3.5: Answer key for the pilot study test

Item	Correct answer	Item	Correct answer
1	C	21*	A
2	C	22	A
3*	A	23	C
4	C	24	C
5	D	25	D
6	B	26	B
7	C	27*	A
8	E	28	C
9*	A	29	C
10	A	30	B
11	B	31	E
12	C	32	C
13	B	33*	A
14	C	34	A
15*	A	35	C
16	B	36	E
17	C	37	C
18	E	38	B
19	D	39	A, B, C, D
20	C	40	B, C, D, E

*These items were open-ended but the correct response was assigned code A and different misconceptions were assigned codes B, C, D, and E.

Appendix 3.6: Final data collection test booklet (English)

Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

Student school Roll No:
School research ID:
Student research ID:

What grade are you in? (tick as appropriate): (a) Grade 4 (b) Grade 5

You are a? (tick as appropriate): (a) Girl (b) Boy

Your year of birth (please write): _____

Have you attended pre-primary (tick as appropriate): (a) Yes (b) No

Do you take private tuition after school? (tick as appropriate): (a) Yes (b) No

Parents' education (if known to you please write) Mother: _____ Father: _____

Parents' profession (if known to you please write) Mother: _____ Father: _____

Family's average monthly expenditure (if known to you please write): _____

(Please answer the following items by putting a tick mark (✓) on the correct answers where applicable. For some items, you will need to write down the answer in the space provided under them. If you want, you can use additional paper for calculations. One example item has been done for you here.)

Example:

Akash has 5 toys. Ria has 3 toys. How many toys do they have together?

- a. 5
- b. 3
- ✓ c. 8
- d. 2
- e. 4
- f. None of the above

1. Jahid had 5 apples. Koli gave him 7 more apples. How many apples did Jahid have in the beginning?
b. 12 b. 7 c. 5 d. 2 e. Not enough information to tell
2. Razia has 15 candles. Daud has 7 candles less than Razia. Which of these number sentences is appropriate for calculating how many candles Daud has?
a. $15 + 7$
b. $7 + 15$
c. $7 - 15$
d. $15 - 7$
e. None of the above

3. Solve this addition problem $2457 + 1319$. Show your work.
4. $522 + 87 = ?$
- 527
 - 609
 - 435
 - 509
 - None of the above
5. Suppose you were asked to add 235 and 54, and you found the result was 289. How would you write the solution down?
- $$\begin{array}{r} 235 \\ 54+ \\ \hline 289 \end{array}$$
 - $$\begin{array}{r} 289 \\ +54 \\ \hline 235 \end{array}$$
 - $$\begin{array}{r} 54+ \\ 234 \\ \hline 289 \end{array}$$
 - $$\begin{array}{r} 235 \\ +54 \\ \hline 289 \end{array}$$
 - None of the above
6. In a cricket match Sujon got 120 runs, Sumon got 85 runs, and Sujoy got 67 runs. How many runs did they score together?
- 262
 - 272
 - 362
 - 372
 - None of the above

7. 3 of Raju's friends gave him some coins. First, Samia gave him 45 coins; Jakaria gave him 80 coins; and then Lisa gave him 25 coins. How many coins did he get from the last one of his friends?

- a. 125
- b. 80
- c. 45
- d. 25
- e. None of the above

8. Latif has 138 plants in his house. Mitu has 112 plants more than him. Which of these number sentences would be appropriate for calculating how many plants Mitu has?

- a. $138 - 12$
- b. $112 - 138$
- c. $140 - 12$
- d. $140 - 138$
- e. None of the above

9. Solve this subtraction problem $4869 - 1344$. Show your work.

10. $328 - 75 = ?$

- a. 253
- b. 403
- c. 263
- d. 273
- e. None of the above

11. Suppose you were asked to subtract 84 from 439 and you found the answer was 355. How would you write the solution down?

a.
$$\begin{array}{r} 84 \\ - 439 \\ \hline 355 \end{array}$$

b.
$$\begin{array}{r} 439 \\ - 84 \\ \hline 355 \end{array}$$

c. 439

$$\begin{array}{r} 84 - \\ \hline 355 \end{array}$$

d. 355

$$\begin{array}{r} + 84 \\ \hline 439 \end{array}$$

e. None of the above

12. Mr. Ajay went to the market with 4250 taka. After doing shopping he had 890 taka left with him. How much money did he spend in the market?

a. 5140

b. 3350

c. 5150

d. 3360

e. None of the above

13. Jamal has more toys than Bakar who has more toys than Rimi but less than Rafi. Which of the following is true?

a. Bakar has more toys than Rafi

b. Rafi has more toys than Rimi

c. Rimi has more toys than Rafi

d. Rafi has more toys than Bakar

e. None of the above

14. Two hundred and twenty-three boys, one hundred and thirty-five girls, and twelve teachers participated in a school trip. Which of these number sentences is correct for finding out the total number of people going on the trip?

a. $223 + 135$

b. $135 + 12 - 223$

c. $223 - 135 - 12$

d. $12 + 135 + 223$

e. None of the above

15. Solve this addition problem: $52758 + 25936$. Show your work.

16. Which of the following is the closest to the result for solving $94 + 33$?
- 110
 - 130
 - 120
 - 140
 - None of the above
17. Suppose you were asked to add 26, 2533, and 325. You found the answer was 2884. How would you write the solution down?
- $$\begin{array}{r} 26 \\ +325 \\ \hline 2884 \end{array}$$
 - $$\begin{array}{r} 2533 \\ 325 \\ - 26 \\ \hline 2884 \end{array}$$
 - $$\begin{array}{r} 2533 \\ 325 \\ + 26 \\ \hline 2884 \end{array}$$
 - $$\begin{array}{r} 2533 \\ 325+ \\ 26 \quad . \\ \hline 2884 \end{array}$$
 - None of the above
18. In a city, there live 23465 male, 2335 female and 11280 children. What is the population of the city?
- 37000
 - 40830
 - 35780
 - 36900
 - None of the above
19. A cricket bat weighs 2 kilograms, a pencil weighs 50 grams, a ruler weighs 100 grams, a television weighs 6 kilograms, and a tomato weighs 30 grams. A cricket bat has the highest length among these but which of these objects weighs more than the cricket bat?

- a. Pencil
 - b. Ruler
 - c. Tomato
 - d. Television
 - e. None of the above
20. Joba has some oranges. Dilu has 28 oranges which is 12 more than Joba. Which of the following number sentences is right for figuring out how many oranges Joba has?
- a. $28 + 12$
 - b. $12 - 28$
 - c. $12 + 28$
 - d. $28 - 12$
 - e. None of the above
21. Solve this subtraction problem $6867 - 2593$. Show your work.
22. Which of the following is the closest to the result for solving $87 - 53$?
- a. 37
 - b. 44
 - c. 27
 - d. 30
 - e. None of the above
23. Suppose you were asked to subtract 129 from 2613 and you found the answer was 2484. How would you write the solution down?
- a.
$$\begin{array}{r} 2631 \\ -129 \\ \hline 2484 \end{array}$$
 - b.
$$\begin{array}{r} -129 \\ 2613 \\ \hline 2884 \end{array}$$
 - c.
$$\begin{array}{r} 2613 \\ -192 \\ \hline 2484 \end{array}$$

d. 2613

$$\begin{array}{r} 129- \\ \hline 2484 \end{array}$$

e. None of the above

24. A school needs 9500 taka for annual sports. The school got 4500 taka as government grant and 2000 taka from school fund. How much more money do they need to collect?

a. 4500

b. 3000

c. 2000

d. 5000

e. None of the above

25. Khulna is to the west from Dhaka, Narsingdi is to the east of Dhaka. Sylhet is further east from Dhaka. Which of these cities is the furthest towards the west?

a. Sylhet

b. Dhaka

c. Narsingdi

d. Khulna

e. Not enough information to tell

26. Rahat has 41 stamps, Rubin has 30 stamps. Lisa has 23 stamps more than what Rahat and Rubin have together. Which of the following number sentences is right for calculating how many stamps Lisa has?

a. $41 + 30 - 23$

b. $41 + 30 + 23$

c. $30 - 23 + 41$

d. $41 - 30 - 23$

e. None of the above

27. Solve this addition problem $40695 + 17582 + 10346$. Show your work.

28. Which of the following is the closest to the result for solving for $2137 + 251$?
- a. 2350
 - b. 2360
 - c. 2420
 - d. 2400
 - e. None of the above
29. A teacher is buying some new furniture for his school. He bought two tables which are 7 feet long and weighs 25 kilograms each. He also bought 5 desks, each of which are 5 feet long and weighs 20 kilograms. The teacher calculated the total weight of the furniture to be 150 kg and decided a method of transportation. Which of the following alternatives show the correct calculation and an appropriate transportation method?
- a. Weight $25+25+25+25+20+20+20= 150$ kilogram; a rickshaw for transportation
 - b. Weight $25+25+25+25+20+20+20= 150$ kilogram; a truck for transportation
 - c. Weight $25+25+20+20+20+20+20 = 150$ kilogram; a truck for transportation
 - d. Weight $25+25+20+20+20+20+20 = 150$ kilogram; a bus for transportation
 - e. None of the above
30. The number of males and females in a town is 42530. Among them, the number of males is 21678; find the number of females.
- a. 19561
 - b. 20852
 - c. 20521
 - d. 19322
 - e. None of the above
31. Water from a swimming pool has some salt in it. Tap water has the least amount of salt. But there is much salt in water from the Atlantic or the Pacific Ocean. Which water contains the most salt?
- a. Tap water
 - b. Swimming pool water
 - c. Water from the Atlantic
 - d. Water from the Pacific
 - e. Not enough information to tell
32. Rubi got 2500 taka from her parents to buy new books. Nusrat got 700 taka less than Rubi and Rima got 300 taka less than Nusrat. Which of the following number sentences is correct for calculating the amount Rima got?

- a. $2500 + 700 + 300$
- b. $2500 - 700 + 300$
- c. $2500 + 700 - 300$
- d. $2500 - 700 - 300$
- e. None of the above

33. Solve this subtraction problem $4912 - 2385 - 1456$. Show your work.

34. Which of the following is the correct answer for solving $7322 - 3980$?

- a. 3342
- b. 4342
- c. 3242
- d. 4242
- e. None of the above

35. The headteacher of a school is planning a victory day celebration in her school and wanted each student to have a small Bangladeshi flag. These flags come in packets of 200s. There are 4 packets of flags in the school. The school has 656 students in total but 126 of them will not be present during the celebration. The headteacher calculated that there is going to be 530 students in the celebration and the school would not need to buy new flags. Which of the following alternatives show the correct calculation and result for the number of students and number of packets of flags required?

- a. $656 + 126 = 530$; 4 packets of flag
- b. $656 - 126 = 530$; 4 packets of flag
- c. $656 - 126 = 530$; 3 packets of flag
- d. $656 - 126 = 530$; 2 packets of flag
- e. None of the above

36. Shihab has 390 taka more than that of Siam and Siam has 470 taka less than that of Shimul. Shimul has 890 taka. How much money does each of Siam and Shihab has?

- a. Siam 420 taka; Shihab 810 taka
- b. Siam 810 taka; Shihab 420 taka
- c. Siam 470 taka; Shihab 390 taka
- d. Siam 860 taka; Shihab 890 taka
- e. None of the above

37. Pia was asked to solve this problem: 'Rashed started for his school from home at 10.40 am and reached the class at 11.30 am. How long did it take for him to reach his class?'

She solved the problem and found the answer was 50. How shall she write down the result?

- a. 50 am
- b. 50 hours
- c. 50 minutes
- d. 0.50 hours
- e. None of the above

38. Leon was asked to solve the following problem: 'Jenny bought 5 candies for 12.5 taka. She gave the shopkeeper a 20-taka note. How much shall she get in change?'

He solved the problem and found the answer was 7.5. How should he write down the answer?

- a. 7 taka
- b. 7 taka and 50 poisa
- c. 7.5 poisa
- d. 7.5
- e. None of the above

(Thank you for your participation)

Appendix 3.7: Final data collection test booklet (Bangla)

Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

শিক্ষার্থীর বিদ্যালয় রোলঃ
বিদ্যালয়ের গবেষণা আইডিঃ
শিক্ষার্থীর গবেষণা আইডিঃ

তুমি কোন শ্রেণীতে পড়? (টিক চিহ্ন দাও) : (ক) ৪র্থ শ্রেণী (খ) ৫ম শ্রেণী

তুমি কী? (টিক চিহ্ন দাও) (ক) ছাত্রী (খ) ছাত্র

তোমার জন্মসাল (পাশের খালি জায়গায় লিখ)

তুমি কি শিশু শ্রেণীতে পড়েছ? (টিক চিহ্ন দাও) (ক) হ্যাঁ (খ) না

তুমি কি স্কুল সময়ের পরে কোন শিক্ষকের কাছে প্রাইভেট পড়? (টিক চিহ্ন দাও) (ক) হ্যাঁ (খ) না

তোমার অভিভাবকদের শিক্ষাগত যোগ্যতা (জানা থাকলে লিখ) : মা বাবা

তোমার অভিভাবকদের পেশা (জানা থাকলে লিখ) : মা বাবা

তোমার পরিবারের গড় মাসিক ব্যয় (জানা থাকলে লিখ) : (টাকা)

(নিচের প্রশ্নগুলোতে যেখানে প্রয়োজ্য সঠিক উত্তরে টিক চিহ্ন দাও। কিছু প্রশ্নের নিচে দেয়া খালি জায়গায় উত্তর লিখতে হবে। তুমি চাইলে হিসেব করার জন্য আলাদা কাগজ ব্যবহার করতে পার। এখানে একটি উদাহরণের সমাধান করা আছে।)

উদাহরণঃ

আকাশের ৫ টি খেলনা আছে। রিয়ার ৩ টি খেলনা আছে। তাদের একত্রে কয়টি খেলনা আছে?

ক) ৫

খ) ৩

✓ গ) ৮

ঘ) ২

ঙ) ৪

০১) জাহিদের কাছে ৫ টা আপেল ছিল। কলি তাকে আরো ৭ টা আপেল দিল। জাহিদের কাছে শুরুতে ক'টা আপেল ছিল?

(ক) ১২ (খ) ৭ (গ) ৫ (ঘ) ২ (ঙ) উত্তর দেবার মত যথেষ্ট তথ্য নেই

০২) রাজিয়ার কাছে ১৫ টি মোমবাতি আছে। দাউদের কাছে রাজিয়ার চেয়ে ৭ টি মোমবাতি কম আছে। দাউদের কাছে কতগুলো মোমবাতি আছে সেটা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

(ক) $১৫ + ৭$

(খ) $৭ + ১৫$

(গ) $৭ - ১৫$

(ঘ) $১৫ - ৭$

(ঙ) উপরের কোনটিই নয়

০৩) নিচের খালি জায়গায় এই যোগ অঙ্কটি করঃ $২৪৫৭ + ১৩১৯$ ।

০৪) $৫২২ + ৮৭ = ?$

- (ক) ৫২৭
- (খ) ৬০৯
- (গ) ৪৩৫
- (ঘ) ৫০৯
- (ঙ) উপরের কোনটিই নয়

০৫) মনে কর তোমাকে বলা হল ২৩৫ আর ৫৪ যোগ করতে। যোগ করে তুমি উত্তর পেলে ২৮৯। এই যোগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

(ক)
$$\begin{array}{r} ২৩৫ \\ ৫৪+ \\ \hline ২৮৯ \end{array}$$

(খ)
$$\begin{array}{r} ২৮৯ \\ +৫৪ \\ \hline ২৩৫ \end{array}$$

(গ)
$$\begin{array}{r} ৫৪+ \\ ২৩৫ \\ \hline ২৮৯ \end{array}$$

(ঘ)
$$\begin{array}{r} ২৩৫ \\ +৫৪ \\ \hline ২৮৯ \end{array}$$

- (ঙ) উপরের কোনটিই নয়

০৬) ক্রিকেট খেলায় সুজন ১২০ রান, সুমন ৮৫ রান এবং সুজয় ৬৭ রান করল। তিনজনে মোট কত রান করল?

- (ক) ২৬২
- (খ) ২৭২
- (গ) ৩৬২
- (ঘ) ৩৭২
- (ঙ) উপরের কোনটিই নয়

০৭) রাজুর তিনজন বন্ধু তাকে কিছু পয়সা দিল। প্রথমে সামিয়া তাকে ৪৫ টি পয়সা দিল, জাকারিয়া তাকে ৮০ টি পয়সা দিল এবং তারপর লিসা তাকে ২৫ টি পয়সা দিল। বন্ধুদের মধ্যে শেষজন রাজুকে কয়টি পয়সা দিয়েছিল?

- (ক) ১২৫
- (খ) ৮০
- (গ) ৪৫
- (ঘ) ২৫
- (ঙ) কোনটিই নয়

০৮) লতিফের বাসায় ১৩৮ টি ফুলগাছ আছে। মিতুর বাসায় লতিফের চেয়ে ১১২ টি ফুলগাছ বেশী আছে। মিতুর বাসায় কতগুলো ফুলগাছ আছে সেটা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $১৩৮ - ১২$
- (খ) $১১২ - ১৩৮$
- (গ) $১৪০ - ১২$
- (ঘ) $১৪০ - ১৩৮$
- (ঙ) উপরের কোনটিই নয়

০৯) নিচের খালি জায়গায় এই বিয়োগ অঙ্কটি করঃ $৪৮৬৯ - ১৩৪৪$ ।

১০) $৩২৮ - ৭৫ = ?$

- (ক) ২৫৩
- (খ) ৪০৩
- (গ) ২৬৩
- (ঘ) ২৭৩
- (ঙ) উপরের কোনটিই নয়

১১) মনে কর, তোমাকে বলা হল ৪৩৯ থেকে ৮৪ বিয়োগ করতে। বিয়োগ করে উত্তর পাওয়া গেল ৩৫৫। এই বিয়োগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

(ক) ৮৪

$$\begin{array}{r} - ৪৩৯ \\ ৩৫৫ \end{array}$$

(খ) ৪৩৯

$$\begin{array}{r} - ৮৪ \\ ৩৫৫ \end{array}$$

(গ) ৪৩৯

$$\begin{array}{r} ৮৪- \\ ৩৫৫ \end{array}$$

(ঘ) ৩৫৫

$$\begin{array}{r} + ৮৪ \\ ৪৩৯ \end{array}$$

(ঙ) উপরের কোনটিই নয়

১২) অজয় বাবু ৪২৫০ টাকা নিয়ে বাজারে গেলেন। বাজার করার পর তার কাছে ৮৯০ টাকা রইল। তিনি কত টাকার বাজার করলেন?

(ক) ৫১৪০

(খ) ৩৩৫০

(গ) ৫১৫০

(ঘ) ৩৩৬০

(ঙ) উপরের কোনটিই নয়

১৩) জামালের বকরের চেয়ে বেশী খেলনা আছে। বকরের রিমির চেয়ে বেশী কিন্তু রাফির চেয়ে কম খেলনা আছে। নিচের কোনটি সত্য?

(ক) বকরের রাফির চেয়ে বেশী খেলনা আছে

(খ) রাফির রিমির চেয়ে বেশী খেলনা আছে

(গ) রিমির রাফির চেয়ে বেশী খেলনা আছে

(ঘ) রাফির বকরের চেয়ে বেশী খেলনা আছে

(ঙ) উপরের কোনটিই নয়

১৪) একটি স্কুল থেকে দুইশ তেইশজন ছাত্র, একশ পঁয়ত্রিশ জন ছাত্রী এবং বারজন শিক্ষক শিক্ষাসফরে যাচ্ছেন। মোট কতজন শিক্ষাসফরে যাচ্ছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $২২৩ + ১৩৫$
- (খ) $১৩৫ + ১২ - ২২৩$
- (গ) $২২৩ - ১৩৫ - ১২$
- (ঘ) $১২ + ১৩৫ + ২২৩$
- (ঙ) উপরের কোনটিই নয়

১৫) নিচের খালি জায়গায় এই যোগ অঙ্কটি করঃ $৫২৭৫৮ + ২৫৯৩৬$ ।

১৬) নিচের কোন বিকল্পটি $৯৪ + ৩৩$ এর যোগফলের সবচেয়ে কাছাকাছি?

- (ক) ১১০
- (খ) ১৩০
- (গ) ১২০
- (ঘ) ১৪০
- (ঙ) উপরের কোনটিই নয়

১৭) মনে কর তোমাকে বলা হল ২৬, ২৫৩৩ আর ৩২৫ যোগ করতে। যোগ করে তুমি উত্তর পেলে ২৮৮৮। এই যোগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

$$\begin{array}{r} (ক) \ ২৪৩৩ \\ \quad \quad ২৬ \\ \quad \quad + ৩২৫ \\ \hline \quad \quad ২৮৮৮ \end{array}$$

$$\begin{array}{r} (খ) \ ২৫৩৩ \\ \quad \quad ৩২৫ \\ \quad \quad - ২৬ \\ \hline \quad \quad ২৮৮৮ \end{array}$$

$$\begin{array}{r} (গ) \ ২৫৩৩ \\ \quad \quad ৩২৫ \\ \quad \quad + ২৬ \\ \hline \quad \quad ২৮৮৮ \end{array}$$

$$\begin{array}{r} (ঘ) \ ২৫৩৩ \\ \quad \quad ৩২৫ + \\ \quad \quad \quad ২৬ \\ \hline \quad \quad ২৮৮৮ \end{array}$$

(ঙ) উপরের কোনটিই নয়

১৮) একটি শহরে ২৩৪৬৫ জন পুরুষ, ২৩৩৫ জন মহিলা ও ১১২৮০ জন শিশু বাস করে। উক্ত শহরের জনসংখ্যা কত?

- (ক) ৩৭০০০
- (খ) ৪০৮৩০
- (গ) ৩৫৭৮০
- (ঘ) ৩৬৯০০

(ঙ) উপরের কোনটিই নয়

১৯) একটি ক্রিকেট ব্যাটের ওজন ২ কিলোগ্রাম, একটি পেসিলের ওজন ৫০ গ্রাম, একটি রুলারের ওজন ১০০ গ্রাম, একটি টেলিভিশনের ওজন ৬ কিলোগ্রাম, এবং একটি টমেটোর ওজন ৩০ গ্রাম। এগুলোর মধ্যে ক্রিকেট ব্যাটটি লম্বায় সবচেয়ে বড় কিন্তু কোন বস্তুটির ওজন ক্রিকেট ব্যাটের চেয়ে বেশী?

- (ক) পেসিলা
- (খ) রুলার
- (গ) টমেটো
- (ঘ) টেলিভিশন
- (ঙ) উপরের কোনটিই নয়

২০) জবার কাছে কিছু কমলালেবু আছে। দিলুর কাছে ২৮ টি কমলালেবু আছে যা জবার চেয়ে ১২ টি বেশী। জবার কাছে ক'টি কমলালেবু আছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $২৮ + ১২$
- (খ) $১২ - ২৮$
- (গ) $১২ + ২৮$
- (ঘ) $২৮ - ১২$
- (ঙ) উপরের কোনটিই নয়

২১) নিচের খালি জায়গায় এই বিয়োগ অঙ্কটি করঃ ৬৮৬৭ - ২৫৯৩ ।

২২) নিচের কোন বিকল্পটি ৮৭ - ৫৩ এর বিয়োগফলের সবচেয়ে কাছাকাছি?

- (ক) ৩৭
- (খ) ৪৪
- (গ) ২৭
- (ঘ) ৩০
- (ঙ) উপরের কোনটিই নয়

২৩) মনে কর, তোমাকে বলা হল ২৬১৩ থেকে ১২৯ বিয়োগ করতে। বিয়োগ করে উত্তর পাওয়া গেল ২৪৮৪। এই বিয়োগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

(ক) ২৬৩১

$$\begin{array}{r} -129 \\ 2613 \\ \hline 2484 \end{array}$$

(খ) - ১২৯

$$\begin{array}{r} 2613 \\ -129 \\ \hline 2484 \end{array}$$

(গ) ২৬১৩

$$\begin{array}{r} -129 \\ 2613 \\ \hline 2484 \end{array}$$

(ঘ) ২৬১৩

$$\begin{array}{r} 129- \\ 2613 \\ \hline 2484 \end{array}$$

(ঙ) উপরের কোনটিই নয়

২৪) বার্ষিক ক্রীড়া প্রতিযোগিতার জন্য ৯৫০০ টাকা প্রয়োজন। সরকারি অনুদান হিসেবে ৪৫০০ টাকা এবং বিদ্যালয়ের তহবিল থেকে ২০০০ টাকা দেয়া হয়েছে। অনুষ্ঠানটি করার জন্য আর কত টাকা যোগাড় করতে হবে?

(ক) ৪৫০০

(খ) ৩০০০

(গ) ২০০০

(ঘ) ৫০০০

(ঙ) উপরের কোনটিই নয়

২৫) খুলনা ঢাকা থেকে পশ্চিম দিকে অবস্থিত, নরসিংদী ঢাকা থেকে পূর্ব দিকে অবস্থিত। সিলেট ঢাকা থেকে আরও পূর্ব দিকে অবস্থিত। এই শহরগুলোর মধ্যে কোনটি সর্বাধিক পশ্চিমে অবস্থিত?

(ক) সিলেট

(খ) ঢাকা

(গ) নরসিংদী

(ঘ) খুলনা

(ঙ) উত্তর দেবার মত যথেষ্ট তথ্য নেই

২৬) রাহাতের কাছে ৪১ টি ডাকটিকিট আছে, রুবিনের কাছে ৩০ টি ডাকটিকিট আছে। লিসার কাছে রাহাত আর রুবিন দু'জনের মিলে যতগুলো ডাকটিকিট তার চেয়ে ২৩ টি ডাকটিকিট বেশী আছে। লিসার কাছে কতটি ডাকটিকিট আছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $৪১ + ৩০ - ২৩$
- (খ) $৪১ + ৩০ + ২৩$
- (গ) $৩০ - ২৩ + ৪১$
- (ঘ) $৪১ - ৩০ - ২৩$
- (ঙ) উপরের কোনটিই নয়

২৭) নিচের খালি জায়গায় এই যোগ অঙ্কটি করঃ $৪০৬৯৫ + ১৭৫৮২ + ১০৩৪৬$ ।

২৮) নিচের কোন বিকল্পটি $২১৩৭ + ২৫১$ এর যোগফলের সবচেয়ে কাছাকাছি?

- (ক) ২৩৫০
- (খ) ২৩৬০
- (গ) ২৪২০
- (ঘ) ২৪০০
- (ঙ) উপরের কোনটিই নয়

২৯) একজন শিক্ষক তার স্কুলের জন্য নতুন আসবাবপত্র কিনেছেন। তিনি দু'টো টেবিল কিনেছেন যেগুলোর প্রতিটির দৈর্ঘ্য ৭ ফুট এবং ওজন ২৫ কিলোগ্রাম। এছাড়াও তিনি ৫ টি বেঞ্চ কিনেছেন যেগুলোর প্রতিটির দৈর্ঘ্য ৫ ফুট এবং ওজন ২০ কিলোগ্রাম। শিক্ষক হিসেব করে দেখলেন আসবাবপত্রগুলোর মোট ওজন ১৫০ কিলোগ্রাম এবং সেগুলোকে স্কুলে নিয়ে যাবার জন্য একটি বাহনের ব্যবস্থা করলেন। নিচের বিকল্পগুলোর মধ্যে কোনটিতে সবচেয়ে সঠিক হিসেব এবং উপযুক্ত পরিবহন মাধ্যম রয়েছে?

- (ক) ওজন $২৫ + ২৫ + ২৫ + ২৫ + ২০ + ২০ + ২০ = ১৫০$ কিলোগ্রাম; পরিবহনের জন্য রিকশা
- (খ) ওজন $২৫ + ২৫ + ২৫ + ২৫ + ২০ + ২০ + ২০ = ১৫০$ কিলোগ্রাম; পরিবহনের জন্য ট্রাক
- (গ) ওজন $২৫ + ২৫ + ২০ + ২০ + ২০ + ২০ + ২০ = ১৫০$ কিলোগ্রাম; পরিবহনের জন্য ট্রাক
- (ঘ) ওজন $২৫ + ২৫ + ২০ + ২০ + ২০ + ২০ + ২০ = ১৫০$ কিলোগ্রাম; পরিবহনের জন্য বাস
- (ঙ) উপরের কোনটিই নয়

৩০) কোন শহরে নারী-পুরুষের সংখ্যা ৪২৫৩০, এদের মধ্যে ২১৬৭৮ জন পুরুষ হলে, নারীর সংখ্যা নির্ণয় কর।

- (ক) ১৯৫৬১
- (খ) ২০৮৫২
- (গ) ২০৫২১
- (ঘ) ১৯৩২২
- (ঙ) উপরের কোনটিই নয়

৩১) সুইমিং পুলের পানিতে খানিকটা লবণ থাকে। কলের পানিতে সবচেয়ে কম লবণ থাকে। তবে আটলান্টিক মহাসাগর বা প্রশান্ত মহাসাগরের পানিতে প্রচুর লবণ থাকে। কোন পানিতে সবচেয়ে বেশী লবণ থাকে?

- (ক) কলের পানিতে
- (খ) সুইমিং পুলের পানিতে
- (গ) আটলান্টিক মহাসাগরের পানিতে
- (ঘ) প্রশান্ত মহাসাগরের পানিতে
- (ঙ) উত্তর দেবার মত যথেষ্ট তথ্য নেই

৩২) নতুন ক্রাসের বই-খাতা কেনার জন্য রুবি বাসা থেকে ২৫০০ টাকা পেয়েছে। নুসরাত রুবির চেয়ে ৭০০ টাকা কম পেয়েছে আর রিমা নুসরাতের চেয়ে ৩০০ টাকা কম পেয়েছে। রিমা কত টাকা পেয়েছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $২৫০০ + ৭০০ + ৩০০$
- (খ) $২৫০০ - ৭০০ + ৩০০$
- (গ) $২৫০০ + ৭০০ - ৩০০$
- (ঘ) $২৫০০ - ৭০০ - ৩০০$
- (ঙ) উপরের কোনটিই নয়

৩৩) নিচের খালি জায়গায় এই বিয়োগ অঙ্কটি করঃ ৪৯১২ - ২৩৮৫ - ১৪৫৬।

৩৪) ৭৩২২ - ৩৯৮০ এর বিয়োগফলের জন্য সঠিক উত্তর নিচের কোনটি?

- (ক) ৩৩৪২
- (খ) ৪৩৪২
- (গ) ৩২৪২
- (ঘ) ৪২৪২
- (ঙ) উপরের কোনটিই নয়

৩৫) একটি স্কুলের প্রধান শিক্ষক স্কুলে বিজয় দিবস উদযাপনের আয়োজন করতে চাইছেন যেখানে সব ছাত্র-ছাত্রীর হাতে বাংলাদেশের একটি করে ছোট পতাকা থাকবে। এই পতাকাগুলো ২০০ টির প্যাকেটে পাওয়া যায়। স্কুলে এমন ৪ টি প্যাকেট আছে। ওই স্কুলে মোট ৬৫৬ জন ছাত্র-ছাত্রী আছে কিন্তু তাদের মধ্যে ১২৬ জন বিজয় দিবসের অনুষ্ঠানে উপস্থিত থাকতে পারবে না। প্রধান শিক্ষক হিসেব করে দেখলেন অনুষ্ঠানে মোট ৫৩০ জন ছাত্র-ছাত্রী থাকবে এবং নতুন করে পতাকাও কিনতে হবে না। নিচের কোন বিকল্পটিতে অনুষ্ঠানে উপস্থিত ছাত্র-ছাত্রী সংখ্যা বের করার জন্য সঠিক গণনা ও উত্তর এবং কত প্যাকেট পতাকা ব্যবহার করতে হবে তার সঠিক উত্তর আছে?

- (ক) $৬৫৬ + ১২৬ = ৫৩০$; ৪ প্যাকেট পতাকা
- (খ) $৬৫৬ - ১২৬ = ৫৩০$; ৪ প্যাকেট পতাকা
- (গ) $৬৫৬ - ১২৬ = ৫৩০$; ৩ প্যাকেট পতাকা
- (ঘ) $৬৫৬ - ১২৬ = ৫৩০$; ২ প্যাকেট পতাকা
- (ঙ) উপরের কোনটিই নয়

৩৬) সিয়াম অপেক্ষা শিহাবের ৩৯০ টাকা বেশি এবং শিমুল অপেক্ষা ৪৭০ টাকা কম আছে। শিমুলের ৮৯০ টাকা আছে। সিয়াম ও শিহাবের প্রত্যেকের কত টাকা আছে?

- (ক) সিয়াম ৪২০ টাকা, শিহাব ৮১০ টাকা
- (খ) সিয়াম ৮১০ টাকা, শিহাব ৪২০ টাকা
- (গ) সিয়াম ৪৭০ টাকা, শিহাব ৩৯০ টাকা
- (ঘ) সিয়াম ৮৬০ টাকা, শিহাব ৮৯০ টাকা
- (ঙ) উপরের কোনটিই নয়

৩৭) পিয়াকে নিচের অঙ্কটি করতে দেয়া হয়েছিল: 'রশেদ কুলে যাবার জন্য বাসা থেকে সকাল ১০.৪০ এ বের হয়ে সকাল ১১.৩০ এ ক্লাসে পৌঁছাল। তার ক্লাসে পৌঁছতে কত সময় লেগেছে?' পিয়া অঙ্কটা করে উত্তর পেল ৫০। উত্তরটা সে কিভাবে লিখবে?

- (ক) সকাল ৫০
- (খ) ৫০ ঘন্টা
- (গ) ৫০ মিনিট
- (ঘ) ০.৫০ ঘন্টা
- (ঙ) উপরের কোনটিই নয়

৩৮) লিয়নকে নিচের অঙ্কটি করতে দেয়া হয়েছিল: 'জেনি ১২.৫ টাকা দিয়ে ৫ টি চকলেট কিনছে। সে দোকানীকে ২০ টাকার একটা নোট দিল। দোকানী তাকে কত ফেরত দেবে?' লিয়ন অঙ্কটা করে উত্তর পেল ৭.৫। উত্তরটা সে কিভাবে লিখবে?

- (ক) ৭ টাকা
- (খ) ৭ টাকা ৫০ পয়সা
- (গ) ৭.৫ পয়সা
- (ঘ) ৭.৫
- (ঙ) উপরের কোনটিই নয়

(সহযোগিতার জন্য তোমাকে ধন্যবাদ)

Appendix 3.8: Answer key for the test administered in the final data collection

Item	Correct answer	Item	Correct answer
1	C	21	A*
2	D	22	A
3*	A	23	E
4	B	24	B
5	D	25	D
6	B	26	B
7	D	27	A*
8	E	28	D
9*	A	29	C
10	A	30	B
11	B	31	E
12	D	32	D
13	B, D	33	A*
14	D	34	A
15*	A	35	C
16	B	36	E
17	C	37	C
18	E	38	B

*These items were open-ended but the correct response was assigned code A and different misconceptions were assigned codes B, C, D, and E.

Appendix 3.9: Expert opinionnaire (English)

Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

Questionnaire for collecting primary teachers' and experts' opinion

Name:

Occupation: E-mail (if available):

Educational Qualification: Contact number (if available):

Level you teach at: Teaching experience:(in years)

In order to explore how students solve arithmetic word problems, the theoretical framework of this research hypothesizes five stages that the students use in solving such problems. Some items representing these five stages have been developed to assess students' skill in them.

Initially, there is a simplified description of the stages here and then some mathematics items have been presented. After reading about the stages, please review the items and choose which stage each item best represents in your opinion from the alternatives given below the items.

- a) *Comprehension of the problem:* this would include reading the problem, analysing and exploring it to have a clear sense of what the problem is about. Successful completion of this stage would give the problem solver a clear about idea the different elements in the problem and the situation described in it.
- b) *Mapping conceptual relations to a numerical mental representation:* this step would involve constructing a quantitative/numerical mental model of the concepts presented in the problem. This can be regarded as mathematization of the problem and real-life and common sense approaches would play a role in this mapping. At this stage the problem solver would have a quantitative/numerical understanding of the initial and the goal state.
- c) *Selecting an appropriate strategy for solution:* this step would include the problem solver exploring his/her strategy repertoire (mental collection of procedures) to identify an appropriate and effective method for solving the problem. If none of the existing strategies are readily suitable for the problem, they might retrofit an existing strategy with concepts they have from other strategies and evaluate its suitability in solving the problem they are trying to solve.

- d) *Performing the calculations on the stored numbers:* in this step the problem solver would execute the plan he/she had made by using the procedure selected. This would mean using the selected strategy/procedure on the quantitative representation of the initial stage to reach the goal stage. Carefulness in carrying out the computation would be an important element of this stage. There may also be low stress methods they try using, make a guess and check.
- e) *Verifying the solution and communicating the results:* this stage would involve reviewing the solution, evaluating its appropriateness and checking for errors. This would provide a cognitive feedback to the problem solver regarding the success/failure of the journey through previous processes. Furthermore, it would include communicating the results in an appropriate format.

Starting from next page, there are some items here. Most of the items are multiple choice questions where students are supposed to select the correct answer. There are also some items where students are to write down the answers. The abovementioned stages are there under each item for you to select. Please select the stage you think the item before it best represents according to your opinion. An example has been done for you here:

Example:

Item: Akash has 5 toys. Ria has 3 toys. How many toys do they have together?

- a. 5
- b. 3
- c. 8
- d. 2
- e. 4

In your opinion which stage does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

(For each item below, please select the stage you think the item best represents)

Jahid had 5 apples. Koli gave him 7 more apples. How many apples did Jahid have in the beginning?

- a. 12 b. 7 c. 5 d. 2 e. Not enough information to tell

01. In your opinion which stage does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Razia has 15 candies. Daud has 7 candies less than Razia. Which of these number sentences would be appropriate for calculating how many candies Daud has?

- a. $15 + 7$
- b. $7 + 15$
- c. $15 - 7$
- d. $7 - 15$
- e. None of the above

02. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Solve this addition $2457 + 1319$. Show your work.

03. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

$522 + 87 = ?$

- a. 527
- b. 609
- c. 435
- d. 509
- e. None of the above

04. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Suppose you were asked to add 235 and 54, and you found the result was 289. How would you write the solution down?

a. 235

$$\begin{array}{r} 54+ \\ \hline 289 \end{array}$$

b. 289

$$\begin{array}{r} +54 \\ \hline 235 \end{array}$$

c. 54+

$$\begin{array}{r} 234 \\ \hline 289 \end{array}$$

d. 235

$$\begin{array}{r} +54 \\ \hline 289 \end{array}$$

e. None of the above

05. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

3 of Raju's friends gave him some coins. First, Samia gave him 45 coins; Jakaria gave him 80 coins; and then Lisa gave him 25 coins. How many coins did he get from the last one of his friends?

- a. 125 b. 80 c. 45 d. 25 e. None of the above

06. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Latif has 138 plants in his house. Mitu has 112 plants more than him. Which of these number sentences would be appropriate for calculating how many plants Mitu has?

- a. $138 - 12$
- b. $112 - 138$
- c. $140 - 12$
- d. $140 - 138$
- e. None of the above

07. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Solve this subtraction problem $4869 - 1344$. Show your work

08. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

$328 - 75 = ?$

- a. 253
- b. 403
- c. 263
- d. 273
- e. None of the above

09. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Suppose you were asked to subtract 84 from 439 and you found the answer was 355.
How would you write the solution down?

a.
$$\begin{array}{r} 84 \\ - 439 \\ \hline 355 \end{array}$$

b.
$$\begin{array}{r} 439 \\ - 84 \\ \hline 355 \end{array}$$

c.
$$\begin{array}{r} 439 \\ 84 - \\ \hline 355 \end{array}$$

d.
$$\begin{array}{r} 355 \\ + 84 \\ \hline 439 \end{array}$$

- e. None of the above

10. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Jamal has more toys than Bakar who has more toys than Rimi but less than Rafi. Which of the following is true?

- a. Bakar has more toys than Rafi
- b. Rafi has more toys than Rimi
- c. Rimi has more toys than Rafi
- d. Rafi has more toys than Bakar
- e. None of the above

11. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Two hundred and twenty-three boys, one hundred and thirty-five girls, and twelve teachers participated in a school trip. Which of these number sentences is correct for finding out the total number of people going on the trip?

- a. $223 + 135$
- b. $135 + 12 - 223$
- c. $223 - 135 - 12$
- d. $12 + 135 + 223$
- e. None of the above

12. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Solve this addition problem: $52758 + 25936$. Show your work.

13. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Which of the following is the closest to the result for solving $94 + 33$?

- a. 110
- b. 130
- c. 120
- d. 140
- e. None of the above

14. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Suppose you were asked to add 26, 2533, and 325. You found the answer was 2884. How would you write the solution down?

a.
$$\begin{array}{r} 2433 \\ 26 \\ +325 \\ \hline 2884 \end{array}$$

b.
$$\begin{array}{r} 2533 \\ 325 \\ -26 \\ \hline 2884 \end{array}$$

c.
$$\begin{array}{r} 2533 \\ 325 \\ +26 \\ \hline 2884 \end{array}$$

d.
$$\begin{array}{r} 2533 \\ 325+ \\ 26 \\ \hline 2884 \end{array}$$

e. None of the above

15. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

A cricket bat weighs 2 kilograms, a pencil weighs 50 grams, a ruler weighs 100 grams, a television weighs 6 kilograms, and a tomato weighs 30 grams. A cricket bat has the highest length among these but which of these objects weighs more than the cricket bat?

- a. Pencil
- b. Ruler
- c. Tomato
- d. Television
- e. None of the above

16. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Joba has some oranges. Dilu has 28 oranges which is 12 more than Joba. Which of the following number sentences is right for figuring out how many oranges Joba has?

- a. $28 + 12$
- b. $12 - 28$
- c. $12 + 28$
- d. $28 - 12$
- e. None of the above

17. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Solve this subtraction $6867 - 2593$. Show your work.

18. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Which of the following is the closest to the result for solving $87 - 53$?

- a. 37
- b. 44
- c. 27
- d. 30
- e. None of the above

19. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Suppose you were asked to subtract 129 from 2613 and you found the answer was 2484. How would you write the solution down?

a.
$$\begin{array}{r} 2631 \\ -129 \\ \hline 2484 \end{array}$$

b.
$$\begin{array}{r} -129 \\ 2613 \\ \hline 2884 \end{array}$$

c.
$$\begin{array}{r} 2613 \\ -192 \\ \hline 2484 \end{array}$$

d.
$$\begin{array}{r} 2613 \\ 129- \\ \hline 2484 \end{array}$$

- e. None of the above

20. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Khulna is to the west from Dhaka, Narsingdi is to the east of Dhaka. Sylhet is further east from Dhaka. Which of these cities is the furthest towards the west?

- a. Sylhet
- b. Dhaka
- c. Narsingdi
- d. Khulna
- e. not enough information to tell

21. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Rahat has 41 stamps, Rubin has 30 stamps. Lisa has 23 stamps more than what Rahat and Rubin have together. Which of the following number sentences is right for calculating how many stamps Lisa has?

- a. $41 + 30 - 23$
- b. $41 + 30 + 23$
- c. $30 - 23 + 41$
- d. $41 - 30 - 23$
- e. None of the above

22. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Solve this addition $40695 + 17582 + 10346$. Show your work.

23. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Which of the following is the closest to the result for solving for $2137 + 251$?

- a. 2350
- b. 2360
- c. 2420
- d. 2400
- e. None of the above

24. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

A teacher is buying some new furniture for his school. He bought two tables which are 7 feet long and weighs 25 kilograms each. He also bought 5 desks, each of which are 5 feet long and weighs 20 kilograms. The teacher calculated the total weight of the furniture to be 150 kg and decided a method of transportation. Which of the following alternatives show the correct calculation and an appropriate transportation method?

- a. Weight $25+25+25+25+20+20+20= 150$ kilogram; a rickshaw for transportation
- b. Weight $25+25+25+25+20+20+20= 150$ kilogram; a truck for transportation
- c. Weight $25+25+20+20+20+20+20 = 150$ kilogram; a truck for transportation
- d. Weight $25+25+20+20+20+20+20 = 150$ kilogram; a bus for transportation
- e. None of the above

25. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Water from a swimming pool has some salt in it. Tap water has the least amount of salt. But there is much salt in water from the Atlantic or the Pacific Ocean. Which water contains the most salt?

- a. Tap water
- b. Swimming pool water
- c. Water from the Atlantic
- d. Water from the Pacific
- e. Not enough information to tell

26. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Rubi got 2500 taka from her parents to buy new books. Nusrat got 700 taka less than Rubi and Rima got 300 taka less than Nusrat. Which of the following number sentences is correct for calculating the amount Rima got?

- a. $2500 + 700 + 300$
- b. $2500 - 700 + 300$
- c. $2500 + 700 - 300$
- d. $2500 - 700 - 300$
- e. None of the above

27. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Solve this subtraction $4912 - 2385 - 1456$. Show your work.

28. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Which of the following is the correct answer for solving $7322 - 3980$?

- a. 3342
- b. 4342
- c. 3242
- d. 4242
- e. None of the above

29. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

The headteacher of a school is planning a victory day celebration in her school and wanted each student to have a small Bangladeshi flag. These flags come in packets of 200s. There are 4 packets of flags in the school. The school has 656 students in total but 126 of them will not be present during the celebration. The headteacher calculated that there is going to be 530 students in the celebration and the school would not need to buy new flags. Which of the following alternatives show the correct calculation and result for the number of students and number of packets of flags required?

- a. $656 + 126 = 530$; 4 packets of flag
- b. $656 - 126 = 530$; 4 packets of flag
- c. $656 - 126 = 530$; 3 packets of flag
- d. $656 - 126 = 530$; 2 packets of flag
- e. None of the above

30. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Pia was asked to solve this problem: 'Rashed started for his school from home at 10.40 am and reached the class at 11.30 am. How long did it take for him to reach his class?'

She solved the problem and found the answer was 50. How shall she write down the result?

- a. 50 am
- b. 50 hours
- c. 50 minutes
- d. 0.50 hours
- e. None of the above

31. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

Leon was asked to solve the following problem: 'Jenny bought 5 candies for 12.5 taka. She gave the shopkeeper a 20-taka note. How much shall she get in change?' He solved the problem and found the answer was 7.5. How should he write down the answer?

- a. 7 taka
- b. 7 taka and 50 poisa
- c. 7.5 poisa
- d. 7.5
- e. None of the above

32. In your opinion which step does the item above best represent?

- a) Comprehension of the problem
- b) Mapping conceptual relations to a numerical representation
- c) Selecting an appropriate strategy for solution
- d) Performing the calculations on the stored numbers
- e) Verifying the solution and communicating the results

(Many thanks for your participation)

Appendix 3.10: Expert opinionnaire (Bangla)

Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving প্রাথমিক স্তরের শিক্ষক ও বিশেষজ্ঞদের মতামত সংগ্রহসংক্রান্ত প্রশ্নমালা

নামঃ
পেশাঃ ই-মেইল (যদি থাকে)
সর্বোচ্চ শিক্ষাগত যোগ্যতাঃ ফোন নম্বরঃ
শিক্ষকতার স্তরঃ শিক্ষকতার অভিজ্ঞতাঃ (বছর)

এই গবেষণার জন্য যে তত্ত্বীয়-কাঠামো ব্যবহার করা হয়েছে তাতে ৫ টি ধাপ রয়েছে। গবেষণাটির অন্যতম মূল ভাবনা এই যে শিক্ষার্থীরা পাটিগণিতের গাণিতিক সমস্যা সমাধানের ক্ষেত্রে এই পাঁচটি ধাপ ব্যবহার করে থাকে। তাই তাদের গাণিতিক সমস্যা সমাধানের দক্ষতা যাচাইয়ের উদ্দেশ্যে এই ৫টি ধাপকে প্রতিনিধিত্ব করে এমন কিছু প্রশ্ন তৈরি করা হয়েছে।

এখানে শুরুতে ধাপগুলোর সংক্ষিপ্ত বর্ণনা দেয়া আছে এবং পরবর্তীতে প্রশ্নগুলো দেয়া আছে। ধাপগুলো সম্পর্কে জানার পরে অনুগ্রহ করে প্রশ্নগুলো পড়ুন এবং আপনার মতে প্রতিটি প্রশ্ন কোন ধাপটিকে সবচেয়ে ভালভাবে প্রতিনিধিত্ব করে তা প্রতিটি প্রশ্নের নিচে দেয়া বিকল্পগুলো থেকে বাছাই করে টিক চিহ্ন দিন।

(ক) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারাঃ

এই ধাপের জন্য প্রশ্নগুলো এমন হবে যেন সেগুলো শিক্ষার্থীদের পঠন দক্ষতা যাচাই করতে পারে। প্রশ্নগুলো শিক্ষার্থীদের গাণিতিক সমস্যা পড়ে বুঝতে পারা এবং জটিলভাবে উপস্থাপিত সমস্যাকে বিশ্লেষণ করে তা থেকে প্রয়োজনীয় তথ্য বের করে আনতে পারার ক্ষমতা যাচাই করবে। অপ্রয়োজনীয় তথ্যকে বাদ দিয়ে দরকারী তথ্য চিনতে পারা যাচাই করাও এ ধাপের প্রশ্নগুলোর উদ্দেশ্য।

(খ) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্কে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারাঃ

বাক্যসম্বলিত গাণিতিক সমস্যা সমাধানের ক্ষেত্রে বাক্য/গল্প থেকে গাণিতিক সংখ্যা এবং তাদের মধ্যকার সম্পর্ক বোঝানোর জন্য প্রয়োজনীয় প্রতীক (+, - ইত্যাদি) নির্ধারণ করতে পারা একটি গুরুত্বপূর্ণ দক্ষতা। এই ধাপের জন্য প্রশ্নগুলো এমন হবে যেন গাণিতিক সমস্যা থেকে সংখ্যা ও প্রতীকগুলোকে নিয়ে তাদের মধ্যকার সম্পর্ক যথার্থভাবে উপস্থাপন করার কাজটিতে শিক্ষার্থীরা কতটা দক্ষ তা যাচাই করা যায়।

(গ) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারাঃ

যোগ বিয়োগ করবার সময় শিক্ষার্থীদেরকে নানান রকমের কৌশল ব্যবহার করতে হয়। প্রাথমিক স্তরে এ ধরনের কৌশলের মধ্যে রয়েছে ‘হাতে রেখে যোগ’, ‘বাম দিকের সংখ্যা থেকে ধার করে বিয়োগ’ ইত্যাদি। একই অংকে একাধিকবার এ ধরনের কৌশল প্রয়োগ করাও উচ্চতর দক্ষতা বলে বিবেচনা করা হয়। এ ধাপের জন্য প্রশ্নগুলো এমন হবে যেন শিক্ষার্থীদের এসব কৌশল ব্যবহারের সামর্থ্য যাচাই করা যায়।

(ঘ) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারাঃ

এই ধাপটি শিক্ষার্থীদের হিসেব করার দক্ষতার সাথে অর্থাৎ তাদের যোগ বিয়োগ করার কুশলতা যাচাই করার সাথে সম্পর্কিত। ধাপটির জন্য প্রশ্নগুলো এমন হবে যেন শিক্ষার্থীদের দুই বা ততোধিক সংখ্যা সঠিকভাবে যোগ বা বিয়োগ করে উত্তর বের করার দক্ষতা যাচাই করা যায়।

(ঙ) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারাঃ

গাণিতিক সমস্যা সঠিকভাবে সমাধানের জন্য প্রচলিত প্রথা মেনে সমাধানটিকে উপস্থাপন করতে পারা, উত্তর লিখতে পারা এবং সমাধান ও উত্তর উভয় যাচাই করতে পারা অত্যন্ত গুরুত্বপূর্ণ দক্ষতা। এই ধাপের জন্য প্রশ্নগুলো এমন হবে যেন শিক্ষার্থীদের গাণিতিক সমাধান যাচাই করতে পারা ও উত্তর সঠিকভাবে উপস্থাপন করতে পারার দক্ষতা পরীক্ষা করা যায়।

পরবর্তী পৃষ্ঠা থেকে কিছু প্রশ্ন দেয়া আছে। অধিকাংশ প্রশ্ন নৈর্ব্যক্তিক ধরণের যেখানে শিক্ষার্থীরা সঠিক উত্তরে টিক চিহ্ন দেবে, কিছু প্রশ্নে শিক্ষার্থীদের উত্তর লিখতে হবে। প্রতিটি প্রশ্নের নিচে আপনার বাছাই করার জন্য উপরে বর্ণিত ধাপগুলো দেয়া আছে। অনুগ্রহ করে প্রতিটি প্রশ্ন পড়ুন এবং আপনার মতে সেটি কোন ধাপকে সবচেয়ে ভালভাবে প্রতিনিধিত্ব করে তা প্রশ্নের নিচে দেয়া ধাপগুলো থেকে বাছাই করে টিক চিহ্ন দিন। উদাহরণ হিসেবে এখানে একটি প্রশ্নের উত্তর করার পদ্ধতি দেয়া আছে।

উদাহরণঃ

আকাশের ৫ টি খেলনা আছে। রিয়ার ৩ টি খেলনা আছে। তাদের মধ্যে দ্বিতীয় জনের কয়টি খেলনা আছে?

- ক) ৫
খ) ৩
গ) ৮
ঘ) ২
ঙ) ৪

আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- ✓ (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
(ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
(iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
(iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
(v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

(অনুগ্রহ করে প্রতিটি প্রশ্ন পড়ুন এবং আপনার মতে সেটি কোন ধাপকে সবচেয়ে ভালভাবে প্রতিনিধিত্ব করে তা প্রশ্নের নিচে দেয়া ধাপগুলো থেকে বাছাই করে টিক চিহ্ন দিন।)

জাহিদের কাছে ৫ টা আপেল ছিল। কলি তাকে আরো ৭ টা আপেল দিল। জাহিদের কাছে শুরুতে ক'টা আপেল ছিল?

- (ক) ১২ (খ) ৭ (গ) ৫ (ঘ) ২ (ঙ) উত্তর দেবার মত যথেষ্ট তথ্য নেই

০১) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

রাজিয়ার কাছে ১৫ টি মোমবাতি আছে। দাউদের কাছে রাজিয়ার চেয়ে ৭ টি মোমবাতি কম আছে। দাউদের কাছে কতগুলো মোমবাতি আছে সেটা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $১৫ + ৭$
(খ) $৭ + ১৫$
(গ) $৭ - ১৫$
(ঘ) $১৫ - ৭$
(ঙ) উপরের কোনটিই নয়

০২) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

নিচের খালি জায়গায় এই যোগ অঙ্কটি করঃ ২৪৫৭ + ১৩১৯।

০৩) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

$৫২২ + ৮৭ = ?$

- (ক) ৫২৭
- (খ) ৬০৯
- (গ) ৪৩৫
- (ঘ) ৫০৯
- (ঙ) উপরের কোনটিই নয়

০৪) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

মনে কর তোমাকে বলা হল ২৩৫ আর ৫৪ যোগ করতে। যোগ করে তুমি উত্তর পেলে ২৮৯। এই যোগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

(ক) ২৩৫

$$\begin{array}{r} ৫৪+ \\ ২৮৯ \end{array}$$

(খ) ২৮৯

$$\begin{array}{r} +৫৪ \\ ২৩৫ \end{array}$$

(গ) ৫৪+

$$\begin{array}{r} ২৩৪ \\ ২৮৯ \end{array}$$

(ঘ) ২৩৫

$$\begin{array}{r} +৫৪ \\ ২৮৯ \end{array}$$

(ঙ) উপরের কোনটিই নয়

০৫) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

রাজুর তিনজন বন্ধু তাকে কিছু পয়সা দিল। প্রথমে সামিয়া তাকে ৪৫ টি পয়সা দিল, জাকারিয়া তাকে ৮০ টি পয়সা দিল এবং তারপর লিসা তাকে ২৫ টি পয়সা দিল। বন্ধুদের মধ্যে শেষজন রাজুকে কয়টি পয়সা দিয়েছিল?

- (ক) ১২৫ (খ) ৮০ (গ) ৪৫ (ঘ) ২৫ (ঙ) কোনটিই নয়

০৬) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

লতিফের বাসায় ১৩৮ টি ফুলগাছ আছে। মিতুর বাসায় লতিফের চেয়ে ১১২ টি ফুলগাছ বেশী আছে। মিতুর বাসায় কতগুলো ফুলগাছ আছে সেটা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) ১৩৮ - ১২
 (খ) ১১২ - ১৩৮
 (গ) ১৪০ - ১২
 (ঘ) ১৪০ - ১৩৮
 (ঙ) উপরের কোনটিই নয়

০৭) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

নিচের খালি জায়গায় এই বিয়োগ অঙ্কটি করঃ ৪৮৬৯ - ১৩৪৪।

০৮) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

৩২৮ - ৭৫ = ?

- (ক) ২৫৩
- (খ) ৪০৩
- (গ) ২৬৩
- (ঘ) ২৭৩
- (ঙ) উপরের কোনটিই নয়

০৯) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

মনে কর, তোমাকে বলা হল ৪৩৯ থেকে ৮৪ বিয়োগ করতে। বিয়োগ করে উত্তর পাওয়া গেল ৩৫৫।
এই বিয়োগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

(ক) ৮৪

- ৪৩৯

৩৫৫

(খ) ৪৩৯

- ৮৪

৩৫৫

(গ) ৪৩৯

৮৪-

৩৫৫

(ঘ) ৩৫৫

+ ৮৪

৪৩৯

(ঙ) উপরের কোনটিই নয়

১০) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্কে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

জামালের বকরের চেয়ে বেশী খেলনা আছে। বকরের রিমির চেয়ে বেশী কিন্তু রাফির চেয়ে কম খেলনা আছে। নিচের কোনটি সত্য?

- (ক) বকরের রাফির চেয়ে বেশী খেলনা আছে
- (খ) রাফির রিমির চেয়ে বেশী খেলনা আছে
- (গ) রিমির রাফির চেয়ে বেশী খেলনা আছে
- (ঘ) রাফির বকরের চেয়ে বেশী খেলনা আছে
- (ঙ) উপরের কোনটিই নয়

১১) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্কে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

একটি স্কুল থেকে দুইশ তেইশজন ছাত্র, একশ পঁয়ত্রিশ জন ছাত্রী এবং বারজন শিক্ষক শিক্ষাসফরে যাচ্ছেন। মোট কতজন শিক্ষাসফরে যাচ্ছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $২২৩ + ১৩৫$
- (খ) $১৩৫ + ১২ - ২২৩$
- (গ) $২২৩ - ১৩৫ - ১২$
- (ঘ) $১২ + ১৩৫ + ২২৩$
- (ঙ) উপরের কোনটিই নয়

১২) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্কে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

নিচের খালি জায়গায় এই যোগ অঙ্কটি করঃ ৫২৭৫৮ + ২৫৯৩৬।

১৩) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

নিচের কোন বিকল্পটি ৯৪ + ৩৩ এর যোগফলের সবচেয়ে কাছাকাছি?

- (ক) ১১০
- (খ) ১৩০
- (গ) ১২০
- (ঘ) ১৪০
- (ঙ) উপরের কোনটিই নয়

১৪) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

মনে কর তোমাকে বলা হল ২৬, ২৫৩৩ আর ৩২৫ যোগ করতে। যোগ করে তুমি উত্তর পেলে ২৮৮৮। এই যোগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

(ক) ২৪৩৩
 ২৬
 +৩২৫
 ২৮৮৮

(খ) ২৫৩৩
 ৩২৫
 - ২৬
 ২৮৮৮

(গ) ২৫৩৩
 ৩২৫
 +২৬
 ২৮৮৮

(ঘ) ২৫৩৩
 ৩২৫+
 ২৬
 ২৮৮৮

(ঙ) উপরের কোনটিই নয়

১৫) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

একটি ক্রিকেট ব্যাটের ওজন ২ কিলোগ্রাম, একটি পেসলের ওজন ৫০ গ্রাম, একটি ব্লসারের ওজন ১০০ গ্রাম, একটি টেলিভিশনের ওজন ৬ কিলোগ্রাম, এবং একটি টমেটোর ওজন ৩০ গ্রাম। এগুলোর মধ্যে ক্রিকেট ব্যাটটি লম্বায় সবচেয়ে বড় কিন্তু কোন বস্তুটির ওজন ক্রিকেট ব্যাটের চেয়ে বেশী?

- (ক) পেসল
- (খ) ব্লসার
- (গ) টমেটো
- (ঘ) টেলিভিশন
- (ঙ) উপরের কোনটিই নয়

১৬) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুবাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্কে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

জ্বার কাছে কিছু কমলালেবু আছে। দিলুর কাছে ২৮ টি কমলালেবু আছে যা জ্বার চেয়ে ১২ টি বেশী। জ্বার কাছে ক'টি কমলালেবু আছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $২৮ + ১২$
- (খ) $১২ - ২৮$
- (গ) $১২ + ২৮$
- (ঘ) $২৮ - ১২$
- (ঙ) উপরের কোনটিই নয়

১৭) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুবাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্কে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

নিচের খালি জায়গায় এই বিয়োগ অঙ্কটি করঃ ৬৮৬৭ - ২৫৯৩ ।

১৮) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

নিচের কোন বিকল্পটি ৮৭ - ৫৩ এর বিয়োগফলের সবচেয়ে কাছাকাছি?

- (ক) ৩৭
- (খ) ৪৪
- (গ) ২৭
- (ঘ) ৩০
- (ঙ) উপরের কোনটিই নয়

১৯) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

মনে কর, তোমাকে বলা হল ২৬১৩ থেকে ১২৯ বিয়োগ করতে। বিয়োগ করে উত্তর পাওয়া গেল ২৪৮৪। এই বিয়োগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

(ক) ২৬৩১

১২৯

২৪৮৪

(খ) - ১২৯

২৬১৩

২৪৮৪

(গ) ২৬১৩

- ১২৯

২৪৮৪

(ঘ) ২৬১৩

১২৯-

২৪৮৪

(ঙ) উপরের কোনটিই নয়

২০) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

(i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা

(ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা

(iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা

(iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা

(v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

খুলনা ঢাকা থেকে পশ্চিম দিকে অবস্থিত, নরসিংদী ঢাকা থেকে পূর্ব দিকে অবস্থিত। সিলেট ঢাকা থেকে আরও পূর্ব দিকে অবস্থিত। এই শহরগুলোর মধ্যে কোনটি সর্বাধিক পশ্চিমে অবস্থিত?

- (ক) সিলেট
- (খ) ঢাকা
- (গ) নরসিংদী
- (ঘ) খুলনা
- (ঙ) উত্তর দেবার মত যথেষ্ট তথ্য নেই

২১) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

রাহাতের কাছে ৪১ টি ডাকটিকিট আছে, রুবিনের কাছে ৩০ টি ডাকটিকিট আছে। লিসার কাছে রাহাত আর রুবিন দু'জনের মিলে যতগুলো ডাকটিকিট তার চেয়ে ২৩ টি ডাকটিকিট বেশী আছে। লিসার কাছে কতটি ডাকটিকিট আছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $81 + 30 - 23$
- (খ) $81 + 30 + 23$
- (গ) $30 - 23 + 81$
- (ঘ) $81 - 30 - 23$
- (ঙ) উপরের কোনটিই নয়

২২) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

নিচের খালি জায়গায় এই যোগ অঙ্কটি করঃ ৪০৬৯৫ + ১৭৫৮২ + ১০৩৪৬।

২৩) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

নিচের কোন বিকল্পটি ২১৩৭ + ২৫১ এর যোগফলের সবচেয়ে কাছাকাছি?

- (ক) ২৩৫০
- (খ) ২৩৬০
- (গ) ২৪২০
- (ঘ) ২৪০০
- (ঙ) উপরের কোনটিই নয়

২৪) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

একজন শিক্ষক তার স্কুলের জন্য নতুন আসবাবপত্র কিনেছেন। তিনি দু'টো টেবিল কিনেছেন যেগুলোর প্রতিটির দৈর্ঘ্য ৭ ফুট এবং ওজন ২৫ কিলোগ্রাম। এছাড়াও তিনি ৫ টি বেঞ্চ কিনেছেন যেগুলোর প্রতিটির দৈর্ঘ্য ৫ ফুট এবং ওজন ২০ কিলোগ্রাম। শিক্ষক হিসেব করে দেখলেন আসবাবপত্রগুলোর মোট ওজন ১৫০ কিলোগ্রাম এবং সেগুলোকে স্কুলে নিয়ে যাবার জন্য একটি বাহনের ব্যবস্থা করলেন। নিচের বিকল্পগুলোর মধ্যে কোনটিতে সবচেয়ে সঠিক হিসেব এবং উপযুক্ত পরিবহন মাধ্যম রয়েছে?

- (ক) ওজন $২৫ + ২৫ + ২৫ + ২৫ + ২০ + ২০ + ২০ = ১৫০$ কিলোগ্রাম; পরিবহনের জন্য রিকশা
- (খ) ওজন $২৫ + ২৫ + ২৫ + ২৫ + ২০ + ২০ + ২০ = ১৫০$ কিলোগ্রাম; পরিবহনের জন্য ট্রাক
- (গ) ওজন $২৫ + ২৫ + ২০ + ২০ + ২০ + ২০ + ২০ = ১৫০$ কিলোগ্রাম; পরিবহনের জন্য ট্রাক
- (ঘ) ওজন $২৫ + ২৫ + ২০ + ২০ + ২০ + ২০ + ২০ = ১৫০$ কিলোগ্রাম; পরিবহনের জন্য বাস
- (ঙ) উপরের কোনটিই নয়

২৫) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্কে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

সুইমিং পুলের পানিতে খানিকটা লবণ থাকে। কলের পানিতে সবচেয়ে কম লবণ থাকে। তবে আটলান্টিক মহাসাগর বা প্রশান্ত মহাসাগরের পানিতে প্রচুর লবণ থাকে। কোন পানিতে সবচেয়ে বেশী লবণ থাকে?

- (ক) কলের পানিতে
- (খ) সুইমিং পুলের পানিতে
- (গ) আটলান্টিক মহাসাগরের পানিতে
- (ঘ) প্রশান্ত মহাসাগরের পানিতে
- (ঙ) উত্তর দেবার মত যথেষ্ট তথ্য নেই

২৬) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্কে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

নতুন ক্রাসের বই-খাতা কেনার জন্য রুবি বাসা থেকে ২৫০০ টাকা পেয়েছে। নুসরাত রুবির চেয়ে ৭০০ টাকা কম পেয়েছে আর রিমা নুসরাতের চেয়ে ৩০০ টাকা কম পেয়েছে। রিমা কত টাকা পেয়েছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

(ক) $২৫০০ + ৭০০ + ৩০০$
 (খ) $২৫০০ - ৭০০ + ৩০০$
 (গ) $২৫০০ + ৭০০ - ৩০০$
 (ঘ) $২৫০০ - ৭০০ - ৩০০$
 (ঙ) উপরের কোনটিই নয়

২৭) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্কে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

নিচের খালি জায়গায় এই বিয়োগ অঙ্কটি করঃ ৪৯১২ - ২৩৮৫ - ১৪৫৬।

২৮) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্কে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

৭৩২২ - ৩৯৮০ এর বিয়োগফলের জন্য সঠিক উত্তর নিচের কোনটি?

- (ক) ৩৩৪২
- (খ) ৪৩৪২
- (গ) ৩২৪২
- (ঘ) ৪২৪২
- (ঙ) উপরের কোনটিই নয়

২৯) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় ধাকা সংখ্যা ও তাদের মধ্যকার সম্পর্কে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় ধাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় ধাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

একটি স্কুলের প্রধান শিক্ষক স্কুলে বিজয় দিবস উদযাপনের আয়োজন করতে চাইছেন যেখানে সব ছাত্র-ছাত্রীর হাতে বাংলাদেশের একটি করে ছোট পতাকা থাকবে। এই পতাকাগুলো ২০০ টির প্যাকেটে পাওয়া যায়। স্কুলে এমন ৪ টি প্যাকেট আছে। ওই স্কুলে মোট ৬৫৬ জন ছাত্র-ছাত্রী আছে কিন্তু তাদের মধ্যে ১২৬ জন বিজয় দিবসের অনুষ্ঠানে উপস্থিত থাকতে পারবে না। প্রধান শিক্ষক হিসেব করে দেখলেন অনুষ্ঠানে মোট ৫৩০ জন ছাত্র-ছাত্রী থাকবে এবং নতুন করে পতাকাও কিনতে হবে না। নিচের কোন বিকল্পটিতে অনুষ্ঠানে উপস্থিত ছাত্র-ছাত্রী সংখ্যা বের করার জন্য সঠিক গণনা ও উত্তর এবং কত প্যাকেট পতাকা ব্যবহার করতে হবে তার সঠিক উত্তর আছে?

- (ক) $৬৫৬ + ১২৬ = ৫৩০$; ৪ প্যাকেট পতাকা
- (খ) $৬৫৬ - ১২৬ = ৫৩০$; ৪ প্যাকেট পতাকা
- (গ) $৬৫৬ - ১২৬ = ৫৩০$; ৩ প্যাকেট পতাকা
- (ঘ) $৬৫৬ - ১২৬ = ৫৩০$; ২ প্যাকেট পতাকা
- (ঙ) উপরের কোনটিই নয়

৩০) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় ধাকা সংখ্যা ও তাদের মধ্যকার সম্পর্কে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় ধাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় ধাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

পিয়াকে নিচের অঙ্কটি করতে দেয়া হয়েছিল: 'রশেদ স্কুলে যাবার জন্য বাসা থেকে সকাল ১০.৪০ এ বের হয়ে সকাল ১১.৩০ এ ক্লাসে পৌছাল। তার ক্লাসে পৌছতে কত সময় লেগেছে?' পিয়া অঙ্কটা করে উত্তর পেল ৫০।

উত্তরটা সে কিভাবে লিখে?

- (ক) সকাল ৫০
- (খ) ৫০ ঘন্টা
- (গ) ৫০ মিনিট
- (ঘ) ০.৫০ ঘন্টা
- (ঙ) উপরের কোনটিই নয়

৩১) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

লিয়নকে নিচের অঙ্কটি করতে দেয়া হয়েছিল: 'জেনি ১২.৫ টাকা দিয়ে ৫ টি চকলেট কিনছে। সে দোকানীকে ২০ টাকার একটা নোট দিল। দোকানী তাকে কত ফেরত দেবে?' লিয়ন অঙ্কটা করে উত্তর পেল ৭.৫। উত্তরটা সে কিভাবে লিখে?

- (ক) ৭ টাকা
- (খ) ৭ টাকা ৫০ পয়সা
- (গ) ৭.৫ পয়সা
- (ঘ) ৭.৫
- (ঙ) উপরের কোনটিই নয়

৩২) আপনার মতে উপরের এই প্রশ্নটি সবচেয়ে ভালভাবে কোন ধাপের প্রতিনিধিত্ব করে?

- (i) গাণিতিক সমস্যা বুঝতে/অনুধাবন করতে পারা
- (ii) গাণিতিক সমস্যায় থাকা সংখ্যা ও তাদের মধ্যকার সম্পর্ককে সাংখ্যিক প্রতীকের মাধ্যমে উপস্থাপন করতে পারা
- (iii) গাণিতিক সমস্যায় থাকা যোগ-বিয়োগ করার জন্য উপযুক্ত কৌশল নির্বাচন করতে পারা
- (iv) গাণিতিক সমস্যায় থাকা সংখ্যাগুলো ব্যবহার করে হিসেব করতে পারা
- (v) গাণিতিক সমস্যা সমাধানের পদ্ধতি যাচাই করতে এবং উত্তর লিখতে পারা

(আপনার সদয় সহযোগিতার জন্য অসংখ্য ধন্যবাদ)

Appendix 3.11: Plain language statement (PLS) and consent form for head teachers

[PLS]



Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

Your school has been invited to participate in research being conducted as a doctoral study under the supervision of Professor Mark Wilson and Professor David Clarke of the Melbourne Graduate School of Education at the University of Melbourne.

What are the aims of the research?

This research aims to develop assessment techniques which will help teachers to identify specific difficulties students face while solving arithmetic word problems. As a consequence, it is hoped that students will get better instruction in areas where they are struggling.

What is your school's involvement in the research?

Students of Grade 4 and Grade 5 in your school will take part in a test which is similar to taking a classroom test in the Mathematics class. They will be asked to respond to some questions on a paper related to arithmetic and to do some calculations. Some demographic information (e.g. age, gender) about them and their scores in Bangla and Mathematics school tests will also be collected. In addition, some selected students will be interviewed. The test should take approximately 60 minutes and the interviews for each student should not take more than 60 minutes. These activities should not cause them anything more than a minor discomfort or inconvenience. In addition, you will be requested to fill out a school information questionnaire where data will be collected about your school, including information on the education and experience of mathematics teachers teaching grades 4 and 5.

Information collected from your school and your students will be used with the highest level of confidentiality. The data will be de-identified and none of the participants will be named at any stage of disseminating the findings. Utmost care will be taken to ensure findings cannot be linked back to individuals. At the end of the study, findings of this research will be shared with the school.

How will confidentiality be protected?

Your school's participation in this study is completely voluntary. Should you wish to withdraw your school, or any of the student's or any of their data that has not been processed, you are completely free to do so. The information gathered during the study will be stored in secured files at the University of Melbourne, under the University's guidelines for the management of research data and records and destroyed five years after publication of written findings.

How do you agree to participate?

If you agree to take part in the study, please indicate that you have read and understood this information by signing the accompanying consent form and returning it to Farhan Azim in scanned format to the email address (fazim@student.unimelb.edu.au).

How do you obtain further information?

Should you require any further information or have any concerns, please do not hesitate to contact Professor Hosne Ara Begum of the University of Dhaka who is supporting as a mentor in this research project. Her email is hosne.dhaka@gmail.com and her phone number is +8801791500218.

This research project has been approved by the Human Research Ethics Committee of The University of Melbourne. If you have any concerns or complaints about the conduct of this research project, which you do not wish to discuss with the research team, you should contact the Manager, Human Research Ethics, Office for Research Ethics and Integrity, University of Melbourne, VIC 3010. Tel: +61 3 8344 2073 or Fax: +61 3 9347 6739 or Email: HumanEthics-complaints@unimelb.edu.au. All complaints will be treated confidentially. In any correspondence please provide the name of the research team or the name or ethics ID number (1648228) of the research project.

Thank you for reading this statement and considering its contents.

Yours sincerely,

Professor Mark Wilson

Melbourne Graduate School of Education
The University of Melbourne
email: wilson.m@unimelb.edu.au

Professor David Clarke

Melbourne Graduate School of Education
The University of Melbourne
email: d.clarke@unimelb.edu.au

Farhan Azim

PhD student
Melbourne Graduate School of Education
The University of Melbourne
email: fazim@student.unimelb.edu.au

[consent form]

Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

Name of Head Teacher:

Name of School:

Name of Investigators: Prof. Mark Wilson, Prof. David Clarke, Farhan Azim

I consent to the participation of my school and students to participate in the above research. I understand the nature of the research and what is expected of my school and students and I agree to the participation of my school on that basis.

I acknowledge that:

- I have received a copy of and read the Plain Language Statement explaining the study;
- I have received an adequate explanation of all likely risks, effects, discomforts or inconvenience arising from participation in the project;
- I understand that participation in the study is voluntary, that I and my students have the right to withdraw at any time, and that we may withdraw any data supplied up to the point of analysis and publication;
- I understand that my students will participate in a test;
- I understand that some demographic information about my students will be collected;
- I understand that my students' scores in Mathematics and Bangla tests in school will be collected;
- I understand that some of my students may be interviewed;
- I understand that some data will be collected regarding my school and teachers;
- I have been informed that the confidentiality of the information I and my students provide will be safeguarded, subject to any legal requirements (e.g., subpoena, freedom of information, mandated reporting), and in the following ways:
 - no names or personal details of individual participants will be revealed in any report of the study, and any contextual details that might reveal their identity will be removed;
 - only the investigators will have access to the data collected in this study; and
 - data will be stored securely in a locked facility in the Melbourne Graduate School of Education, and disposed-off securely five years after the completion of the written findings of the research.

Signature of Head Teacher:..... **Date:**

Appendix 3.12: School background questionnaire (English)

Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

School Research ID:

School Information

(Please select the appropriate option from the alternatives or give the requested information for the following questions)

1. School Name:
2. School Address:
Sub-district: District:
3. School's geographic location: (a) Rural (b) Semi-urban (c) Urban
4. School type: (a) Government (b) High-school attached
(c) Registered (d) Private (e) Other (please write).....
5. Total number of students in primary (please write): Boys: Girls:
6. Total number of students in grade 4 (please write):.....Boys: Girls:
7. Number of students who are repeating grade 4 (if any):
8. Total number of students in grade 5 (please write): Boys: Girls:
9. Number of students who are repeating grade 5 (if any):
10. Total number of teachers in primary (please write): Male: Female:
11. Length of each Mathematics class in grade 4: minutes
12. Length of each Mathematics class in grade 5: minutes
13. Number of mathematics classes in a typical week in grade 4:
14. Number of mathematics classes in a typical week in grade 5:
15. Yearly fees for grade 4: (taka)
16. Yearly fees for grade 5:(taka)

Mathematics Teachers' information:

17. Details of Grade 4 Mathematics Teacher:

- Gender: (a) Male (b) Female
- Highest academic qualification (please write):
- Has the teacher studied in science stream at any stage? (tick as appropriate)
(a) Yes (b) No
- Which stream did the teacher study in his/ her highest qualification? (tick as appropriate)

(a) Science (b) Business studies (c) Humanities

➤ Has the teacher received any training for teaching mathematics? (tick as appropriate)

(a) Yes (b) No

➤ Details of training received on teaching mathematics (if applicable, please write):

.....

➤ Overall teaching experience (please write):(years) (months)

➤ Mathematics teaching experience in grade 4 (please write): (years) (months)

18. Details of grade 5 Mathematics Teacher:

➤ Gender: (a) Male (b) Female

➤ Highest academic qualification (please write):

➤ Has the teacher studied in science stream at any stage? (tick as appropriate)

(a) Yes (b) No

➤ Which stream did the teacher study in his/ her highest qualification? (tick as appropriate)

(a) Science (b) Business studies (c) Humanities

➤ Has the teacher received any training for teaching mathematics? (tick as appropriate)

(a) Yes (b) No

➤ Details of training received on teaching mathematics (if applicable, please write):

.....

➤ Overall teaching experience (please write):(years) (months)

➤ Mathematics teaching experience in grade 4 (please write): (years) (months)

(Thank you for your help in this research)

Appendix 3.13: School background questionnaire (Bangla)

**Assessment of Mathematical Problem Solving:
Addition and Subtraction in Arithmetic Word Problem
Solving**

School Research ID:

বিদ্যালয় সংক্রান্ত তথ্যাবলী

(মেগাবোর্ন বিশ্ববিদ্যালয় পরিচালিত এই গবেষণায় অংশগ্রহণে সম্মতি প্রদানের জন্য আপনাকে ধন্যবাদ। এখানে আপনার দেয়া বিদ্যালয় সংক্রান্ত তথ্য কঠোর গোপনীয়তার সাথে কেবলমাত্র এই গবেষণার জন্য ব্যবহার করা হবে এবং গবেষণার কোন পর্যায়ে বিদ্যালয় বা শিক্ষককে সনাক্ত করা যায় এমন তথ্য প্রকাশ করা হবে না।

অনুগ্রহ করে নিচের প্রশ্নগুলোতে যেখানে প্রযোজ্য টিক চিহ্ন দিন অথবা খালি জায়গায় প্রযোজ্য তথ্য লিখুন। প্রয়োজনে তথ্য প্রদান করার জন্য অতিরিক্ত কাগজ ব্যবহার করতে পারেন।)

১. বিদ্যালয়ের নামঃ
২. বিদ্যালয়ের ঠিকানা-
উপজেলাঃ জেলাঃ
৩. বিদ্যালয়ের ভৌগোলিক অবস্থান (টিক চিহ্ন দিন)ঃ (ক) গ্রাম (খ) শহরতলী (গ) শহর
৪. বিদ্যালয়ের ধরণ (টিক চিহ্ন দিন)ঃ (ক) সরকারী (খ) উচ্চবিদ্যালয় (গ) রেজিস্টার্ড (ঘ) প্রাইভেট
(ঙ) অন্যান্য (লিখুন)
৫. প্রাথমিক বিদ্যালয়ে মোট শিক্ষার্থী সংখ্যা (লিখুন)ঃ
৬. ৪র্থ শ্রেণীতে মোট শিক্ষার্থী সংখ্যা (লিখুন)ঃ ছাত্রঃ ছাত্রীঃ
৭. ৪র্থ শ্রেণীতে দ্বিতীয়বার পড়ছে এমন শিক্ষার্থী সংখ্যা (যদি থাকে)ঃ
৮. ৫ম শ্রেণীতে মোট শিক্ষার্থী সংখ্যা (লিখুন)ঃ ছাত্রঃ ছাত্রীঃ
৯. ৫ম শ্রেণীতে দ্বিতীয়বার পড়ছে এমন শিক্ষার্থী সংখ্যা (যদি থাকে)ঃ
১০. প্রাথমিক বিদ্যালয়ে মোট শিক্ষক সংখ্যা (লিখুন)ঃ পুরুষঃ মহিলাঃ
১১. ৪র্থ শ্রেণীতে প্রতিটি গণিত পিরিয়ডের সময়কালঃ (মিনিট)
১২. ৫ম শ্রেণীতে প্রতিটি গণিত পিরিয়ডের সময়কালঃ (মিনিট)
১৩. আপনার বিদ্যালয়ের ৪র্থ শ্রেণীতে ক্লাস রুটিন অনুযায়ী সপ্তাহে কতটি গণিত ক্লাস থাকে?
১৪. আপনার বিদ্যালয়ের ৫ম শ্রেণীতে ক্লাস রুটিন অনুযায়ী সপ্তাহে কতটি গণিত ক্লাস থাকে?
১৫. বিদ্যালয়ে ৪র্থ শ্রেণীর বাৎসরিক বেতনঃ (টাকা)

১৬. বিদ্যালয়ে ৫ম শ্রেণীর বাৎসরিক বেতনঃ (টাকা)

গণিত বিষয়ের শিক্ষক সম্পর্কিত তথ্যাবলী

১৭. ৪র্থ শ্রেণীর গণিত শিক্ষকের তথ্যঃ

- জন্ম তারিখ (টিক চিহ্ন দিন)ঃ (ক) পুরুষ (খ) মহিলা
- শিক্ষাগত যোগ্যতা (আপনার সর্বোচ্চ ডিগ্রী লিখুন)ঃ
- শিক্ষা জীবনের কোন পর্যায়ে বিজ্ঞান বিভাগে পড়েছেন কি? (টিক চিহ্ন দিন) (ক) হ্যাঁ (খ) না
- সর্বোচ্চ শিক্ষাগত যোগ্যতার বিষয়ের গ্রুপ/ধারা : (ক) বিজ্ঞান (খ) ব্যবসায় শিক্ষা (গ) মানবিক
- গণিত শিক্ষণ বিষয়ে কোন বিশেষ প্রশিক্ষণ পেয়েছেন কি? (টিক চিহ্ন দিন) (ক) হ্যাঁ (খ) না
- গণিত শিক্ষণ সংক্রান্ত প্রশিক্ষণ (যদি থাকে লিখুন)ঃ
- শিক্ষকতার মোট অভিজ্ঞতা (লিখুন)ঃ (বছর) (মাস)
- ৪র্থ শ্রেণীতে গণিত পাঠদানের অভিজ্ঞতা (লিখুন)ঃ (বছর) (মাস)

১৮. ৫ম শ্রেণীর গণিত শিক্ষকের তথ্যঃ

- জন্ম তারিখ (টিক চিহ্ন দিন)ঃ (ক) পুরুষ (খ) মহিলা
- শিক্ষাগত যোগ্যতা (আপনার সর্বোচ্চ ডিগ্রী লিখুন)ঃ
- শিক্ষা জীবনের কোন পর্যায়ে বিজ্ঞান বিভাগে পড়েছেন কি? (টিক চিহ্ন দিন) (ক) হ্যাঁ (খ) না
- সর্বোচ্চ শিক্ষাগত যোগ্যতার বিষয়ের গ্রুপ/ধারা : (ক) বিজ্ঞান (খ) ব্যবসায় শিক্ষা (গ) মানবিক
- গণিত শিক্ষণ বিষয়ে কোন বিশেষ প্রশিক্ষণ পেয়েছেন কি? (টিক চিহ্ন দিন) (ক) হ্যাঁ (খ) না
- গণিত শিক্ষণ সংক্রান্ত প্রশিক্ষণ (যদি থাকে লিখুন)ঃ
- শিক্ষকতার মোট অভিজ্ঞতা (লিখুন)ঃ (বছর) (মাস)
- ৫ম শ্রেণীতে গণিত পাঠদানের অভিজ্ঞতা (লিখুন)ঃ (বছর) (মাস)

(গবেষণাটিতে আপনার সহযোগিতার জন্য ধন্যবাদ)

Appendix 3.14: Revision of items during translation

Item 1: initially item said, ‘Jahid had 5 apples, Koli gave him 7 apples. How many apples did Jahid have in the beginning?’. This was changed for ease of translation, and to make it more sensible to students. The changed version reads, “Jahid had 5 apples, Koli gave him 7 more apples. How many apples did Jahid have in the beginning?”. Experts suggested this would be easier for the students to understand.
Item 2: For this item, the question part ‘which of these number sentences would be appropriate for calculating how many candles Daud has?’ was translated to Bangla as “which of these number sentences would be correct for calculating how many candles Daud has?”. This was done because the Bangla word for correct is easier for students to understand compared to the Bangla word for appropriate and it is unambiguous/clearer.
Item 3: In the English version the item is phrased as “Solve this addition problem $2457 + 1319$. Show your work.”. But to make it similar to the style that Bangladeshi students are more familiar with, it was rephrased in Bangla as. “Solve this addition problem in the space provided below: $2457 + 1319$.”
Item 5: The item is phrased in English as “Suppose you were asked to add 235 and 54, and found the result was 289. How would you write the solution down?”. To make it easier to understand for the Bangladeshi students the translation was changed to “Suppose you were asked to add 235 and 54. After adding them you got the answer 289. Which of the following alternative ways would you choose to write down the solution?”.
Item 7: The item is phrased in English as “First, Samia gave him 45 coins; then Jakaria gave him 80 coins; and finally Lisa gave him 25 coins.”. In the translation it was phrased as “First, Samia gave him 45 coins, Jakaria gave him 80 coins; and then Lisa gave him 25 coins.” This was done to make it an easier read in Bangla for the students.
Item 8: In English this item is, “Latif has 138 plants in his house. Mitu has 112 plants more than him. Which of these number sentences would be appropriate for calculating how many plants Mitu has?” It was translated in Bangla as “Latif has 138 <i>flower plants</i> in his house. Mitu has 112 <i>flower plants</i> more in <i>her house</i>. Which of

these number sentences would be correct for calculating how many <i>flower</i> plants Mitu has?
Item 9: In the English version the item is phrased as “Solve this subtraction $4869 - 1344$. Show your work.”. But to make it similar to the style that Bangladeshi students are more familiar with, it was rephrased in Bangla as. “Solve this subtraction in the space provided below: $4869 - 1344$.”
Item 11: The item is phrased in English as “Suppose you were asked to subtract 84 from 439 and you found the answer was 355. How would you write the solution down?”. To make it easier to understand for the Bangladeshi students the translation was changed to “Suppose you were asked to subtract 84 from 439. You found the answer is 355. Which of the following alternative ways would you choose to write down the solution?”
Item 13: In English the item is, “Jamal has more toys than Bakar who has more toys than Rimi but less than Rafi. Which of the following is true?”. While translating to Bangla it was rephrased as “Jamal has more toys than Bakar. Bakar has more toys than Rimi but Less toys than Rafi. Which of the following is true?”
Item 15: In the English version the item is phrased as “Solve this addition problem: $52758 + 25936$ ”. But to make it similar to the style that Bangladeshi students are more familiar with, it was rephrased in Bangla as. “Solve this addition problem in the space provided below: $52758 + 25936$.”
Item 17: The item is phrased in English as “Suppose you were asked to add 26, 2533, and 325. You found the answer was 2884. How would you write the solution down?”. To make it easier to understand for the Bangladeshi students the translation was rephrased in Bangla as “Suppose you were asked to add 26, 2533, and 325. You found the answer was 2884. Which of the following alternative ways would you choose to write down the solution?”
Item 21: The English version has this item as, “Solve this subtraction $6867 - 2593$. Show your work.” It was rephrased while translating to Bangla as. “Solve this subtraction problem in the space provided below: $6867 - 2593$.”
Item 23: The item was phrased in the English version as “Suppose you were asked to subtract 129 from 2613 and you found the answer was 2484. How would you

write the solution down?” To make it easier for Bangladeshi students to understand, the Bangla translation was rephrased as, “Suppose you were asked to subtract 129 from 2613. You found the answer after subtraction 2484. Which of the following alternative ways would you choose to write down the solution?”.

Item 27: The item is phrased as “Solve this addition $40695 + 17582 + 10346$. Show your work” in the English version. In the Bangla translation it was rephrased as “Solve this addition problem in the space provided below: $40695 + 17582 + 10346$.” This is a format/phrasing students are more familiar with.

Item 29: Phrasing of the question sentence in the item was changed during translation to Bangla. The English version reads, “Which of the following alternatives show the most appropriate solution?” This was rephrased in the Bangla version as, “Which of the following alternatives presents the most correct calculation and appropriate transport medium?” This change was made to make the item more easily readable in Bangla.

Item 33: In the English version the item is phrased as, “Solve this subtraction $4912 - 2385 - 1456$. Show your work.” To fit it better with the Bangladeshi students’ expected style, it was rephrased as, “Solve this subtraction problem in the space provided below: $4912 - 2385 - 1456$ ”.

Appendix 3.15: Plain language statement (PLS) and consent form for parents (English)

[PLS]



Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

Your child has been invited to participate in research being conducted as a doctoral study under the supervision of Professor Mark Wilson and Prof. David Clarke of the Melbourne Graduate School of Education at the University of Melbourne.

What are the aims of the research?

This research aims to develop assessment techniques which will help teachers to identify specific difficulties students face while solving arithmetic word problems. As a consequence, it is hoped that students will get better instruction in areas where they are struggling.

What is your child's involvement in the research?

Your child will take part in a test which is similar to taking a classroom test in the Mathematics class. She/he will be asked to respond to some questions on a paper related to arithmetic and to do some calculations. In addition, her/his scores for school tests of Bangla and Mathematics and some demographic information (e.g. age, gender, parent's profession, parent's education, family's monthly income, etc.) about her/him will be collected. S/he also might be interviewed regarding how s/he has answered the arithmetic questions. The test will take approximately 60 minutes and if s/he is interviewed that should be an additional 60 minutes at most. These activities should not cause him/her anything more than a minor discomfort or inconvenience.

The education of your child will not be altered in any way by you agreeing or not agreeing for him/her to be part of this research. No information about your child's identity will be supplied to us. At the end of the study, findings of this research will be shared with the school.

How will confidentiality be protected?

Your child's participation in this study is completely voluntary. Should you wish to withdraw your child or any of his/her data that has not been processed, you are completely free to do so. We will protect your child's anonymity and privacy to the fullest possible extent. Confidentiality will be protected as any information that could identify your child will not be supplied to us or used in any stage of the research. The information gathered during the study will be stored in secured files at the University of Melbourne, under the University's guidelines for the management of research data and records, and destroyed five years after publication of written findings.

How do you agree to let your child participate?

If you agree for your child to take part in the study, please indicate that you have read and understood this information by signing the accompanying consent form and returning it to your child's school. After this is signed, you may still withdraw your child from the research at any time before the analysis and publication of results.

How do you obtain further information?

Should you require any further information or have any concerns, please do not hesitate to contact Professor Hosne Ara Begum of University of Dhaka who is working as a mentor in this research project. Her email is hosne.dhaka@gmail.com and her phone number is +8801791500218.

This research project has been approved by the Human Research Ethics Committee of The University of Melbourne. If you have any concerns or complaints about the conduct of this research project, which you do not wish to discuss with the research team, you should contact the Manager, Human Research Ethics, Office for Research Ethics and Integrity, University of Melbourne, VIC 3010. Tel: +61 3 8344 2073 or Fax: +61 3 9347 6739 or Email: HumanEthics-complaints@unimelb.edu.au. All complaints will be treated confidentially. In any correspondence please provide the name of the research team or the name or ethics ID number (1648228) of the research project.

Thank you for reading this statement and considering its contents.

Yours sincerely,

Professor Mark Wilson

Melbourne Graduate School of Education
The University of Melbourne
email: wilson.m@unimelb.edu.au

Professor David Clarke

Melbourne Graduate School of Education
The University of Melbourne
email: d.clarke@unimelb.edu.au

Farhan Azim

PhD Student
Melbourne Graduate School of Education
The University of Melbourne
email: fazim@student.unimelb.edu.au

[Consent form]

Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

Name of student: Student ID (school).....

Name of Parent/Guardian:

Name of School:

Name of Investigators: Professor Mark Wilson, Professor David Clarke, Farhan Azim

I consent for my child to participate in the above research. I understand the nature of the research and what is expected of my child.

I acknowledge that:

- I have received a copy of and read the Plain Language Statement explaining the study;
- I have received an adequate explanation of all likely risks, effects, discomforts or inconvenience arising from participation in the project;
- I understand that participation in the study is voluntary, that I have the right to withdraw my consent for my child to participate in the study at any time, and that I may withdraw any data supplied before the analysis and publication of results;
- I understand that my child will participate in a test;
- I understand that my child may be interviewed;
- I understand that my child's Mathematics and Bangla scores from school tests will be collected
- I understand that some demographic information will be collected regarding my child;
- I have been informed that the confidentiality of the information provided by my child will be safeguarded, subject to any legal requirements (e.g., subpoena, freedom of information, mandated reporting), and in the following ways:
 - no names or personal details of individual participants will be revealed in any report of the study, and any contextual details that might reveal their identity will be removed;
 - only the investigators and the teacher conducting the test and interviews will have access to the data collected in this study; and
 - data will be stored securely in a locked facility in the Assessment Research Centre, Melbourne Graduate School of Education, and destroyed five years after the completion of the written findings of the research.

Signature of Parent/Guardian:..... **Date:**

Appendix 3.16: Plain language statement (PLS) and consent form for parents (Bangla)

[Bangla PLS]



Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

শুভেচ্ছা! মেলবোর্ন বিশ্ববিদ্যালয়ের গ্রাজুয়েট স্কুল অফ এডুকেশনের অধীনস্থ মূল্যায়ন গবেষণা কেন্দ্র পরিচালিত একটি গবেষণায় অংশগ্রহণের জন্য আপনার সন্তানকে আমন্ত্রণ জানানো হয়েছে। গবেষণাটি অধ্যাপক মার্ক উইলসন ও অধ্যাপক ডেভিড ক্লার্কের তত্ত্বাবধানে পরিচালিত হচ্ছে।

গবেষণাটির উদ্দেশ্য কী?

এ গবেষণাটির উদ্দেশ্য পাটিগণিতের সমস্যা সমাধানের সময় শিক্ষার্থীরা যেসব অসুবিধার সম্মুখীন হয় শিক্ষকদের জন্য তা নির্ণয়ের উপযোগী মূল্যায়ন পদ্ধতির উন্নতিসাধন করা। আশা করা যায় এর ফলে ভবিষ্যতে শিক্ষার্থীদের সুনির্দিষ্ট অসুবিধা অনুধাবন এবং সে অনুযায়ী উপযুক্ত শিক্ষন পরিকল্পনা প্রণয়ন সহজতর হবে।

এ গবেষণায় আপনার সন্তানের ভূমিকা কী?

আপনার সন্তান গণিত বিষয়ে নেয়া ক্লাস টেস্টের মত একটি পরীক্ষায় অংশগ্রহণ করবে। তাকে পাটিগণিত সংক্রান্ত কিছু প্রশ্নের উত্তর দিতে আর কিছু হিসাব করতে বলা হবে। সেই সাথে তার এবং তার পরিবার সম্পর্কে কিছু জনতাত্ত্বিক (যেমন: বয়স, জেন্ডার, অভিভাবকের পেশা, অভিভাবকের শিক্ষাগত যোগ্যতা, পরিবারের মাসিক আয়, ইত্যাদি) তথ্যও সংগ্রহ করা হবে এবং বিদ্যালয়ের পরীক্ষায় বাংলা ও গণিত বিষয়ে প্রাপ্ত তার ফলাফলও সংগ্রহ করা হবে। সে কীভাবে পাটিগণিত সংক্রান্ত প্রশ্নগুলোর উত্তর করেছে সে ব্যাপারে তার একটি সাক্ষাৎকারও নেয়া হতে পারে। সব মিলিয়ে এতে দুই ঘন্টার বেশী সময় লাগবে না এবং সামান্য অস্বস্তি বা অস্বাচ্ছন্দ্য বোধ ছাড়া তার আর কোন অসুবিধা হবে না।

এ গবেষণায় অংশগ্রহণ করা বা না করার কারণে তার বিদ্যালয়ে শিক্ষালাভের সুবিধার কোন পরিবর্তন হবে না। আপনার সন্তানের পরিচয় সংক্রান্ত কোন তথ্য আমাদেরকে সরবরাহ করা হবে না। গবেষণাটি শেষ হলে এর ফলাফল স্কুলকে জানানো হবে।

গোপনীয়তা কীভাবে রক্ষা করা হবে?

এ গবেষণায় আপনার সন্তানের অংশগ্রহণ সম্পূর্ণরূপে ঐচ্ছিক। আপনি চাইলে তথ্য প্রক্রিয়াজাত হবার পূর্বের যেকোন ধাপে আপনার সন্তান বা তার দেয়া কোন তথ্য প্রত্যাহার করে নিতে পারেন। আমরা আপনার সন্তানের পরিচয়ের গোপনীয়তা সর্বোচ্চ গুরুত্ব দিয়ে রক্ষা করব। আপনার সন্তানকে সনাক্ত করা যায় এমন কোন তথ্য আমাদেরকে সরবরাহ করা হবে না এবং গবেষণার কোন স্তরেও ব্যবহার করা হবে না, কাজেই গোপনীয়তা রক্ষার ব্যাপারে আপনি নিশ্চিত থাকতে পারেন। গবেষণায় সংগৃহীত সকল তথ্য মেলবোর্ন বিশ্ববিদ্যালয়ের তথ্য ব্যবস্থাপনা সংক্রান্ত নীতিমালা অনুযায়ী বিশ্ববিদ্যালয়ে নিরাপদে সংরক্ষণ করা হবে এবং লিখিত ফলাফল প্রকাশিত হওয়ার পাঁচ বছর পরে বিনষ্ট করে ফেলা হবে।

কীভাবে আপনি আপনার সন্তানের অংশগ্রহণের ব্যাপারে সম্মতি দেবেন?

যদি এ গবেষণায় আপনার সন্তানের অংশগ্রহণের ব্যাপারে আপনার সম্মতি থাকে তাহলে আপনি এ সংক্রান্ত তথ্যগুলো পড়েছেন এবং বুঝেছেন তা নিশ্চিত করে সংযুক্ত সম্মতিপত্রটিতে অনুগ্রহ করে সাক্ষর করুন এবং আপনার সন্তানের

HREC: 1648228.1; Date: 13/08/20; Version: Bangla PLS and Consent form for Parent/Guardian (Test & Interview)_v2

Melbourne Graduate School of Education
The University of Melbourne Victoria 3010 Australia
T: +61 3 8344 8285 F: +61 3 8344 8529 W: www.edfac.unimelb.edu.au

বিদ্যালয়ে পাঠিয়ে দিন। এটি সাক্ষর করার পরও গবেষণাটির তথ্য বিশ্লেষণ ও ফলাফল প্রকাশিত হওয়ার পূর্ব পর্যন্ত আপনি আপনার সন্তানকে গবেষণাটি থেকে প্রত্যাহার করতে পারবেন।

কীভাবে আপনি এই গবেষণাটি সম্পর্কে অধিকতর তথ্য পেতে পারেন?

আপনার যদি এই গবেষণাটি নিয়ে আরও কোন প্রশ্ন বা কোন উদ্বেগ থাকে তাহলে আপনি আপনার সন্তানের বিদ্যালয়ের প্রধান শিক্ষককে জানাতে পারেন। তিনি আপনাকে গবেষকদের সাথে যোগাযোগের ব্যাপারে সাহায্য করবেন। প্রয়োজনে আপনি ঢাকা বিশ্ববিদ্যালয়ের অধ্যাপক হোসনে আরা বেগমের সাথে যোগাযোগ করতে পারেন। তিনি এ গবেষণায় একজন মেন্টর হিসেবে থাকতে সদয় সম্মতি দিয়েছেন। তাঁর ই-মেইল: hosne.dhaka@gmail.com।

এই গবেষণাটি মেলবোর্ন বিশ্ববিদ্যালয়ের হিউম্যান রিসার্চ এথিক্স কমিটির অনুমোদনে পরিচালিত হচ্ছে। আপনার যদি গবেষণাটির পরিচালনা সংক্রান্ত কোন বিষয়ে কোন উদ্বেগ থাকে যা আপনি গবেষকদের কারও সাথে আলোচনা করতে চান না, তাহলে এই ঠিকানায় যোগাযোগ করতে পারেন : Manager, Human Research Ethics, Office for Research Ethics and Integrity, University of Melbourne, VIC 3010. Tel: +61 3 8344 2073; Fax: +61 3 9347 6739; Email: HumanEthics-complaints@unimelb.edu.au. সকল অনুসন্ধান/অভিযোগ গোপনীয়তার সাথে বিবেচনা করা হবে। যেকোন যোগাযোগে অনুগ্রহ করে গবেষণাদলের নাম অথবা এথিক্স আইডি নম্বর উল্লেখ করুন।

এই বিবরণটি পড়ার জন্য এবং এর বিষয়বস্তু বিবেচনা করার জন্য আপনাকে ধন্যবাদ।

অধ্যাপক মার্ক উইলসন
মেলবোর্ন গ্রাজুয়েট স্কুল অফ এডুকেশন
মেলবোর্ন বিশ্ববিদ্যালয়

ই-মেইল: wilson.m@unimelb.edu.au

অধ্যাপক ডেভিড ক্লার্ক
মেলবোর্ন গ্রাজুয়েট স্কুল অফ এডুকেশন
মেলবোর্ন বিশ্ববিদ্যালয়

ই-মেইল: d.clarke@unimelb.edu.au

ফারহান আজিম
পি.এইচ.ডি গবেষক
মেলবোর্ন গ্রাজুয়েট স্কুল অফ এডুকেশন
মেলবোর্ন বিশ্ববিদ্যালয়

ই-মেইল: fazim@student.unimelb.edu.au

Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

শিক্ষার্থীর নামঃ শ্রেণীঃ শিক্ষার্থী রোল নং (বিদ্যালয়ের)

অভিভাবকের নামঃ

বিদ্যালয়ের নামঃ

গবেষকদের নামঃ অধ্যাপক মার্ক উইলসন, অধ্যাপক ডেভিড ক্লার্ক, ফারহান আজিম

আমি ওপরের গবেষণাটিতে আমার সন্তানের অংশগ্রহণের ব্যাপারে সম্মতি দিচ্ছি। আমি গবেষণাটির প্রকৃতি এবং আমার সন্তানকে কী করতে হবে সে সম্পর্কে অবগত হয়েছি।

আমি স্বীকার করছি যেঃ

- আমি সাধারণ ভাষায় গবেষণাটির একটি বর্ণনা পেয়েছি এবং পড়েছি;
- গবেষণাটিতে অংশগ্রহণের ফলে সৃষ্ট সম্ভাব্য ঝুঁকি, প্রতিক্রিয়া, অস্বস্তি অথবা অসুবিধা সম্পর্কে আমি উপযুক্ত ব্যাখ্যা পেয়েছি;
- আমি জানি যে গবেষণাটিতে অংশগ্রহণ শেচ্ছাপ্রণোদিত। গবেষণাটি থেকে আমার সন্তানকে অথবা তার দেয়া তথ্যকে গবেষণাটির তথ্য বিশ্লেষণ এবং ফলাফল প্রকাশের পূর্বে যেকোন সময় প্রত্যাহার করে নেবার স্বাধীনতা আমার আছে;
- আমি বুঝতে পেরেছি যে আমার সন্তান একটি পরীক্ষায় অংশগ্রহণ করবে;
- আমি বুঝতে পেরেছি যে আমার সন্তানের সাক্ষাৎকার নেয়া হবে;
- আমি বুঝতে পেরেছি যে আমার সন্তানের স্কুল পরীক্ষায় বাংলা ও গণিত বিষয়ের নম্বর সংগ্রহ করা হবে;
- আমি বুঝতে পেরেছি যে আমার সন্তান সম্পর্কে কিছু জনতাত্ত্বিক তথ্য সংগ্রহ করা হবে;
- আমাকে জানানো হয়েছে যে তথ্য প্রদানের আইনানুগ বাধ্যবাধকতার ব্যাতিত আমার সন্তানের কাছ থেকে সংগৃহীত তথ্যের গোপনীয়তা রক্ষা করা হবে নিম্নবর্ণিত উপায়েঃ
 - গবেষণাটিতে অংশগ্রহণকারীদের নাম বা ব্যক্তিগত তথ্য এককভাবে কোন রিপোর্টে উপস্থাপন করা হবে না। কোন প্রাসঙ্গিক তথ্য থেকে পরিচয় প্রকাশের সম্ভাবনা থাকলে সেটিও অপসারণ করা হবে;
 - কেবলমাত্র গবেষকগণ এবং পরীক্ষা ও সাক্ষাৎকার গ্রহণকারী শিক্ষক সংগৃহীত তথ্য দেখবেন;
 - গবেষণায় সংগৃহীত তথ্য ৫ বছরের জন্য মেলবোর্ন বিশ্ববিদ্যালয়ের মূল্যায়ন গবেষণা কেন্দ্রে সংরক্ষণ করা হবে এবং লিখিত ফলাফল প্রকাশের ৫ বছর পরে বিনষ্ট করে ফেলা হবে।

অভিভাবকের স্বাক্ষরঃ তারিখঃ.....

HREC: 1648228.1; Date: 13/08/20; Version: Bangla PLS and Consent form for Parent/Guardian (Test & Interview)_v2

Melbourne Graduate School of Education
The University of Melbourne Victoria 3010 Australia
T: +61 3 8344 8285 F: +61 3 8344 8529 W: www.edfac.unimelb.edu.au

Appendix 3.17: Plain language statement (PLS) for students (English)

Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

Hello! My name is Farhan Azim. I am a student at the University of Melbourne. I am doing a project to find out how children your age solve mathematical problems. When I finish my project it will be part of my degree, called a "PhD". My teachers, Prof. Mark Wilson and Prof. David Clarke are helping me with my project. They are called my "supervisors". We all work in the Melbourne Graduate School of Education.

Your school's head teacher has given me permission to send you this letter to tell you a bit about my project. Once you have read the letter you can decide if you would like to take part. You should talk to your parents about the project too.

If you decide to be part of the project, I am going to ask you to read and answer some questions about yourself and your family and solve some mathematical problems. You and some other students from your class who are taking part would go into a classroom for about an hour to read the questions and write answers. A teacher will be there to explain about the questions and collect the answers at the end. In addition, you may be interviewed about how you answered the questions. If you want to stop doing the questions, you can at any time you like. If you don't know an answer, or you don't want to answer a question, that's fine too.

Only my supervisors, the teacher taking your test and I will see your answers, so please don't worry that your school teachers might look at them. The project will have nothing to do with your school report or your grades. You don't even have to write your name on the test booklet.

After the project is over, I will lock all the test booklets away safely at the University for 5 years. I have to do this because it is a University rule. After that my supervisors will destroy them.

Remember, you don't have to take part unless you want to. If you have any questions you should talk to your teacher or a parent. If they don't know the answer to your question, they can contact me, or my supervisors, or the University for you.

Professor Mark Wilson

Melbourne Graduate School of Education
The University of Melbourne
email: wilson.m@unimelb.edu.au

Professor David Clarke

Melbourne Graduate School of Education
The University of Melbourne
email: d.clarke@unimelb.edu.au

Farhan Azim

PhD Student
Melbourne Graduate School of Education
The University of Melbourne
email: fazim@student.unimelb.edu.au

Appendix 3.18: Plain language statement (PLS) for students (Bangla)



Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

শুভেচ্ছা! আমার নাম ফারহান আজিম। আমি মেলবোর্ন বিশ্ববিদ্যালয়ের একজন শিক্ষার্থী। আমি একটা গবেষণা করছি যার উদ্দেশ্য তোমার বয়সী শিক্ষার্থীরা কীভাবে গাণিতিক সমস্যার সমাধান করে তা বোঝা। গবেষণাটি সাফল্যের সাথে শেষ করলে এটি আমার একটি ডিগ্রীর অংশ হবে যার নাম পি.এইচ.ডি। আমার শিক্ষকদ্বয় - অধ্যাপক মার্ক উইলসন এবং অধ্যাপক ডেভিড ক্লার্ক আমাকে এই গবেষণায় সাহায্য করছেন। তারা এই গবেষণায় আমার 'সুপারভাইজর'। আমরা সবাই মেলবোর্ন বিশ্ববিদ্যালয়ে কাজ করি।

তোমাদের বিদ্যালয়ের প্রধান শিক্ষক মহোদয় তোমাকে এই চিঠিটি লিখতে অনুমতি দিয়েছেন যাতে আমি গবেষণাটি সম্পর্কে তোমাকে জানাতে পারি। চিঠিটি পড়ার পর তুমি গবেষণাটিতে অংশ নেবে কি না সে ব্যাপারে সিদ্ধান্ত নিতে পার। এ ব্যাপারে তুমি তোমার অভিভাবকের সাথেও কথা বলতে পার।

তুমি যদি গবেষণাটিতে অংশগ্রহণ করতে রাজি থাকো তাহলে আমি তোমাকে কিছু গাণিতিক সমস্যার সমাধান করতে আর তোমার এবং তোমার পরিবার সম্পর্কে কিছু প্রশ্নের উত্তর দিতে অনুরোধ করব। তোমার ক্লাসের অন্যান্য শিক্ষার্থীদের সাথে তুমি একটা ক্লাসরুমে গিয়ে এই প্রশ্নগুলোর উত্তর দেবে। সব মিলিয়ে এতে প্রায় এক ঘন্টা সময় লাগবে। প্রশ্নের প্রয়োজনীয় ব্যাখ্যা দিতে আর তোমার লেখা শেষে উত্তরগুলো সংগ্রহ করতে একজন শিক্ষক সেখানে উপস্থিত থাকবেন। প্রশ্নগুলোর উত্তর দেবার পাশাপাশি তোমার একটি সাক্ষাৎকার নেয়া হতে পারে যেখানে তুমি কীভাবে প্রশ্নগুলোর উত্তর দিয়েছ সে ব্যাপারে তোমার সাথে কথা বলা হবে। যেকোন সময়ে উত্তর করা বন্ধ করতে চাইলে তুমি তা করতে পার। তুমি যদি কোন প্রশ্নের উত্তর না জান, বা যদি কোন প্রশ্নের উত্তর না দিতে চাও তাতেও সমস্যা নেই।

কেবলমাত্র তোমার পরীক্ষা গ্রহণকারী শিক্ষক, আমি আর আমার গবেষণার সুপারভাইজররা তোমার উত্তরগুলো দেখবে। কাজেই তোমার বিদ্যালয়ের শিক্ষক উত্তর দেখবেন এমনটা ভেবে ভয় পেয়ো না।

HREC: 1648228.1; Date: 13/08/20; Version: PLS for students_Bangla

Melbourne Graduate School of Education
The University of Melbourne Victoria 3010 Australia
T: +61 3 8344 8285 F: +61 3 8344 8529 W: www.edfac.unimelb.edu.au

এই গবেষণা এবং তোমার উত্তর দেবার সাথে তোমার বিদ্যালয়ের ফলাফলের কোন সম্পর্ক নেই।
এমনকি তোমার উত্তরের সাথে তোমার নামও লিখতে হবে না।
আমার বিশ্ববিদ্যালয়ের নিয়ম অনুযায়ী গবেষণাটি শেষ হবার পর তোমাদের উত্তর লেখা কাগজগুলো ৫ বছরের জন্য বিশ্ববিদ্যালয়ে নিরাপদে সংরক্ষণ করা হবে। এ সময় পার হলে আমার সুপারভাইজররা কাগজগুলো বিনষ্ট করে দেবেন।

মনে রেখ, এ গবেষণায় তুমি অংশগ্রহণ করতে না চাইলে তোমাকে অংশগ্রহণ করতে হবে না। তোমার কোন প্রশ্ন থাকলে তোমার শিক্ষক বা অভিভাবকের সাথে কথা বলতে পার। তাদের যদি প্রশ্নের উত্তর জানা নাও থাকে, তারা আমার সাথে বা আমার সুপারভাইজর অথবা আমার বিশ্ববিদ্যালয়ের অফিসের সাথে যোগাযোগ করতে পারবেন।

তোমাকে ধন্যবাদ।

অধ্যাপক মার্ক উইলসন
মেলবোর্ন গ্রাজুয়েট স্কুল অফ এডুকেশন
মেলবোর্ন বিশ্ববিদ্যালয়

ই-মেইল: wilson.m@unimelb.edu.au

ফারহান আজিম
পি.এইচ.ডি গবেষক
মেলবোর্ন গ্রাজুয়েট স্কুল অফ এডুকেশন
মেলবোর্ন বিশ্ববিদ্যালয়

ই-মেইল: fazim@student.unimelb.edu.au

অধ্যাপক ডেভিড ক্লার্ক
মেলবোর্ন গ্রাজুয়েট স্কুল অফ এডুকেশন
মেলবোর্ন বিশ্ববিদ্যালয়

ই-মেইল: d.clarke@unimelb.edu.au

Appendix 3.19: Assent form for students (English)

Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

Assent form for students

My name is _____

I have read the plain language statement and understand what I am being asked to do.

(Please tick one of the boxes)

☐

Yes I would like to participate in the test and interview conducted by the researchers from the University of Melbourne

☐

No I do not want to participate in the test and interview conducted by the researchers from the University of Melbourne

And sign here: _____

HREC: 1648228.1; Date: 13/08/20; Version: Assent form for Students_Pilot&FinalData Collection_Eng

Melbourne Graduate School of Education
The University of Melbourne Victoria 3010 Australia
T: +61 3 8344 8285 F: +61 3 8344 8529 W: www.edfac.unimelb.edu.au

Appendix 3.20: Assent form for students (Bangla)

Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

শিক্ষার্থী সম্মতিপত্র

আমার নাম _____

আমি এ গবেষণার বর্ণনাটি পড়েছি এবং আমাকে কী করতে হবে তা বুঝতে পেরেছি।

(গবেষণায় অংশগ্রহণের ব্যাপারে তোমার সম্মতি বা অসম্মতির ওপর নির্ভর করে নিচের যেকোন একটি
ঘরে টিক চিহ্ন দাও।)

☐

হ্যাঁ আমি মেলবোর্ন বিশ্ববিদ্যালয়ের গবেষকদেরকে এ
গবেষণাটিতে অংশগ্রহণ করতে রাজি আছি।

☐

না আমি মেলবোর্ন বিশ্ববিদ্যালয়ের গবেষকদেরকে এ
গবেষণাটিতে অংশগ্রহণ করতে রাজি নই।

HREC: 1648228.1; Date: 13/08/20; Version: Assent form for Students_Pilot&FinalDataCollection_Bangla

Melbourne Graduate School of Education
The University of Melbourne Victoria 3010 Australia
T: +61 3 8344 8285 F: +61 3 8344 8529 W: www.edfac.unimelb.edu.au

*Note: The Bangla version was used with the students. English version was for
the supervisors and for the ethics approval.*

Appendix 3.21: Students' cognitive interview fieldnote template for the pilot study (English)

Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

Student school S/N:
School research ID:
Student research ID:

Example: Akash has 5 toys. Ria has 3 toys. How many toys do they have together?

- a. 5
- b. 3
- ✓ c. 8
- d. 2
- e. 4
- f. none of the above

Initial notes: Grade _____, Class Roll _____, Male/Female

(The interviewer explains how the think-aloud is going to take place. He uses the example item above to demonstrate how to articulate thinking while answering an item. Then the respondent is requested to start answering the test items. Respondents are given blank papers in case they want to make notes or small calculations. There is space under each item for the interviewer to record the responses. Additional paper can also be used for note taking.)

1. Jahid had 5 apples, and then Koli gave him 7 apples. How many apples did Jahid have in the beginning?

- a. 12 b. 7 c. 5 d. 2 e. Not enough information to tell

Item 1 Notes:

2. Razia has 15 candles. Daud has 7 candles less than Razia. Which of these number sentences would be appropriate for calculating how many candles Daud has?

- a. $15 + 7$
- b. $7 + 15$
- c. $15 - 7$
- d. $7 - 15$
- e. None of the above

Item 2 Notes:

3. Solve this addition problem $24 + 31$. Show your work.

Item 3 Notes:

4. $22 + 7 = ?$

- a. 27
- b. 24
- c. 29
- d. 30
- e. None of the above

Item 4 Notes:

5. Suppose you were asked to add 235 and 54, and found the result was 289. How would you write the solution down?

a.
$$\begin{array}{r} 235 \\ 54+ \\ \hline 289 \end{array}$$

b.
$$\begin{array}{r} 289 \\ +54 \\ \hline 235 \end{array}$$

c.
$$\begin{array}{r} 54+ \\ 234 \\ \hline 289 \end{array}$$

d.
$$\begin{array}{r} 235 \\ +54 \\ \hline 289 \end{array}$$

e. None of the above

Item 5 Notes:

6. In a cricket match Sujon got 120 runs, Sumon got 85 runs, and Sujoy got 67 runs. How many runs did they score together?

- a. 205
- b. 272
- c. 152
- d. 372
- e. None of the above

Item 6 Notes:

7. Three of Raju's friends gave him some coins. First, Samia gave him 45 coins; then Jakaria gave him 80 coins; and finally Lisa gave him 25 coins. How many coins did he get from the last one of his friends?

a. 125 b. 80 c. 25 d. 45 e. None of the above

Item 7 Notes:

8. Latif has 138 plants in his house. Mitu has 112 plants more than him. Which of these number sentences would be appropriate for calculating how many plants Mitu has?

a. $138 - 12$
b. $112 - 138$
c. $140 - 12$
d. $140 - 138$
e. None of the above

Item 8 Notes:

9. Solve this subtraction problem $26 - 12$. Show your work.

Item 9 Notes:

10. $23 - 8 = ?$

- a. 15
- b. 31
- c. 17
- d. 16
- e. None of the above

Item 10 Notes:

11. Suppose you were asked to subtract 84 from 439 and you found the answer was 355. How would you write the solution down?

a.
$$\begin{array}{r} 84 \\ - 439 \\ \hline 355 \end{array}$$

b.
$$\begin{array}{r} 439 \\ - 84 \\ \hline 355 \end{array}$$

c.
$$\begin{array}{r} 439 \\ 84 - \\ \hline 355 \end{array}$$

d.
$$\begin{array}{r} 355 \\ + 84 \\ \hline 439 \end{array}$$

- e. None of the above

Item 11 Notes:

- 12. Mr. Ajay went to the market with 4250 taka. After doing shopping he had 890 taka left with him. How much money did he spend in the market?**
- a. 5140
 - b. 890
 - c. 3360
 - d. 4210
 - e. None of the above

Item 12 Notes:

- 13. Jamal has 5 toys, Bakar has 8 toys. Which of the following is true?**
- a. Jamal has more toys than Bakar
 - b. Bakar has more toys than Jamal
 - c. Bakar has less toys than Jamal
 - d. Jamal and Bakar have equal numbers of toys
 - e. None of the above

Item 13 Notes:

- 14. Two hundred and twenty-three boys, one hundred and thirty-five girls, and twelve teachers participated in a school trip. Which of these number sentences is correct to find out the total number of people going on the trip?**
- a. $223 + 135$
 - b. $135 + 12 - 223$
 - c. $12 + 135 + 223$
 - d. $223 - 135 - 12$
 - e. None of the above

Item 14 Notes:

15. Solve this addition problem $58 + 27$. Show your work.

Item 15 Notes:

16. Which of the following is the closest to the result for solving $94 + 33$?

- a. 110
- b. 130
- c. 120
- d. 140
- e. None of the above

Item 16 Notes:

17. Suppose you were asked to add 26, 2533, and 325. You found the answer was 2884. How would you write the solution down?

a.
$$\begin{array}{r} 2433 \\ 26 \\ +325 \\ \hline 2884 \end{array}$$

b.
$$\begin{array}{r} 2533 \\ 325 \\ - 26 \\ \hline 2884 \end{array}$$

c.
$$\begin{array}{r} 2533 \\ 325 \\ + 26 \\ \hline 2884 \end{array}$$

d.
$$\begin{array}{r} 2533 \\ 325+ \\ \hline 26 \end{array}$$

2884

e. None of the above

Item 17 Notes:

18. In a city, there live 23465 male, 2335 female and 11289 children. What is the number of population of the city?

- a. 37080
- b. 40830
- c. 35780
- d. 36900
- e. None of the above

Item 18 Notes:

19. A cricket bat weighs 2 kilograms, a pencil weighs 50 grams, a ruler weighs 100 grams, a television weighs 6 kilograms, and a tomato weighs 30 grams. A cricket bat has the highest length among these but which of these objects weighs more than the cricket bat?

- a. Pencil
- b. Ruler
- c. Tomato
- d. Television
- e. None of the above

Item 19 Notes:

20. Joba has some oranges. Dilu has 28 oranges which is 12 more than Joba. Which of the following number sentences is right for figuring out how many oranges Joba has?

- a. $28 + 12$
- b. $12 + 28$
- c. $28 - 12$
- d. $40 - 12$
- e. None of the above

Item 20 Notes:

21. Solve this subtraction problem $63 - 19$. Show your work.

Item 21 Notes:

22. Which of the following is the closest to the result for solving $87 - 53$?

- a. 37
- b. 45
- c. 27
- d. 30
- e. None of the above

Item 22 Notes:

23. Suppose you were asked to subtract 129 from 2613 and you found the answer was 2484. How would you write the solution down?

a.
$$\begin{array}{r} 2631 \\ -129 \\ \hline 2484 \end{array}$$

b.
$$\begin{array}{r} -129 \\ 2613 \\ \hline 2884 \end{array}$$

c.
$$\begin{array}{r} 2613 \\ -129 \\ \hline 2484 \end{array}$$

d.
$$\begin{array}{r} 2613 \\ 129- \\ \hline 2484 \end{array}$$

e. None of the above

Item 23 Notes:

24. A school needs 9500 taka for annual sports. The school got 4500 taka as government grant and 2000 taka from school fund. How much more money do they need to collect?

- a. 4500
- b. 2000
- c. 3000
- d. 5000
- e. None of the above

Item 24 Notes:

25. Khulna is to the west from Dhaka, Narsingdi is to the east of Dhaka. Sylhet is further east from Dhaka. Which of these cities is furthest towards the west?

- a. Sylhet
- b. Dhaka
- c. Narsingdi
- d. Khulna
- e. Not enough information to tell

Item 25 Notes:

26. Rahat has 41 stamps, Rubin has 30 stamps. Lisa has 23 stamps more than what Rahat and Rubin has together. Which of the following number sentences is right for calculating how many stamps Lisa has?

- a. $41 + 30 - 23$
- b. $41 + 30 + 23$
- c. $30 - 23 + 41$
- d. $41 - 30 - 23$
- e. None of the above

Item 26 Notes:

27. Solve this addition problem $1039 + 142 + 14$. Show your work.

Item 27 Notes:

28. Which of the following is the closest to the result for solving for $2137 + 251$?

- a. 2350
- b. 2360
- c. 2400
- d. 2420
- e. None of the above

Item 28 Notes:

29. A teacher is buying some new furniture for his school. He bought 7 desks. 2 of these are 7 feet long and weighs 25 kilo each. The other 5 are 5 feet long and weighs 20 kilo each. The teacher calculated the total weight of the desks to be 150 kg and arranged for a transportation. Which of the following alternatives show the accurate calculation and the appropriate transportation?

- a. Weight $25+25+25+25+20+20+20= 150$ kg; a rickshaw for transportation
- b. Weight $25+25+25+25+20+20+20= 150$ kg; a rickshaw van for transportation
- c. Weight $25+25+20+20+20+20+20 = 150$ kg; a truck for transportation
- d. Weight $25+25+20+20+20+20+20 = 150$ kg; a rickshaw van for transportation
- e. Weight $25+25+20+20+20+20+20 = 150$ kg; a bus for transportation

Item 29 Notes:

30. The number of males and females in a town is 42530. Among them, the number of males is 21678; find the number of females.

- a. 19561
- b. 20852
- c. 20521
- d. 19322
- e. None of the above

Item 30 Notes:

31. Water from a swimming pool has some salt in it. Tap water has the least salt. But there is much salt in water from the Atlantic or the Pacific Ocean. Which water contains the most salt?

- a. Tap water
- b. Swimming pool water
- c. Water from Atlantic
- d. Water from the Pacific
- e. Not enough information to tell

Item 31 Notes:

32. Rubi got 2500 taka from her parents to buy new books. Nusrat got 700 taka less than Rubi and Rima got 300 taka less than Nusrat. Which of the following number sentences is correct for calculating the amount Rima got?

- a. $2500 + 700 + 300$
- b. $2500 - 700 + 300$
- c. $2500 - 700 - 300$
- d. $700 + 300 - 2500$

e. None of the above

Item 32 Notes:

33. Solve this subtraction problem $1269 - 233 - 12$. Show your work.

Item 33 Notes:

34. Which of the following is the correct answer for solving $1245 - 118$?

- a. 1127
- b. 1117
- c. 1137
- d. 1023
- e. None of the above

Item 34 Notes:

35. The headteacher of a school is planning a victory day celebration in her school and wanted each student to have a small Bangladeshi flag. These flags come in packets of 200s. The school has 656 students in total but 126 of them will not be present. The headteacher calculated there will be 530 students in the celebration and the school has 4 packets of flag. Which of the following alternatives show the correct calculation and result?

- a. $656 + 126 = 530$; 4 packets of flag
- b. $656 - 126 = 530$; 4 packets of flag
- c. $656 - 126 = 530$; 3 packets of flag
- d. $656 - 126 = 530$; 2 packets of flag
- e. $656 + 126 = 530$; 3 packets of flag

Item 35 Notes:

36. Shihab has 390 taka more than that of Siam and Siam has 470 taka less than that of Shimul. Shimul has 890 taka. How much money does each of Siam and Shihab has?

- a. Siam 470 taka; Shihab 390 taka
- b. Siam 810 taka; Shihab 420 taka
- c. Siam 420 taka; Shiahb 810 taka
- d. Siam 860 taka; Shihab 890 taka
- e. None of the above

Item 36 Notes:

37. Pia was asked to solve this problem:

'Rashed started for his school from home at 10.40 am and reached the class at 11.30 am. How long did it take for him to reach his class?'

She solved the problem and found the answer was 50. How shall she write down the result?

- a. 50 am
- b. 50 hours
- c. 50 minutes
- d. 0.50 hours
- e. None of the above

Item 37 Notes:

38. Leon was asked to solve the following problem:

‘Jenny bought 5 candies for 12.5 taka. She gave the shopkeeper a 20-taka note. How much shall she get in change?’

He solved the problem and found the answer was 7.5. How should he write down the answer?

- a. 7 taka
- b. 7 taka and 50 poisa

Item 38 Notes:

- c. 7.5 poisa
- d. 7.5
- e. None of the above

39. Think about how you would solve for this addition: 25+32. Which of the following most closely matches your thinking process?

- a. you do column addition

$$\begin{array}{r} 25 \\ +32 \\ \hline 57 \end{array}$$

- b. you remember $25 + 32 = 57$
- c. first you do $20 + 30 = 50$, then $5+2=7$, and then $50+7=57$

- d. first you do $32 + 8 = 40$, then $25 - 8 = 17$; and then $40 + 17 = 57$
- e. you are not sure how to solve this problem
- f. you are going to solve it in some other way. Please write down here:

Item 39 Notes:

40. Think about how you would solve for this problem: $14 - 8$. Which of the following most closely matches your thinking process?

- a. you guess/you don't know how to solve this problem
- b. you remember $14 - 8 = 6$
- c. you count up from 8 till you get to 14 (8, 9, 10,...,14) to get the answer 6
- d. you remember $8 = 4 + 4$, and $14 - 4 = 10$, and then $10 - 4 = 6$
- e. you will do columnar subtraction

$$\begin{array}{r} 14 \\ -8 \\ \hline 6 \end{array}$$

- f. some other method – please write an explanation of your method:

Item 40 Notes:

Additional Questions:

1. Did you find any item of this test problematic while answering? Which item(s)? What problem did you face?

Additional Q1 Notes:

2. What strategies do you use while doing addition and subtraction? Are there certain strategies that you use more?

Additional Q2 Notes:

3. How do you decide which strategy/strategies to use?

Additional Q3 Notes:

4. How would you do (this addition) $25+37$? Tell me step-by-step.

Additional Q4 Notes:

5. How would you do (this subtraction) $54-28$? Tell me step-by-step.

Additional Q5 Notes:

6. After solving a mathematical problem, how do you check whether the answer you have got is correct?

Additional Q6 Notes:

7. What mistakes do you think you make the most while solving arithmetic word problems?

Additional Q7 Notes:

Appendix 3.22: Students' cognitive interview fieldnote template for the pilot study (Bangla)

Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

শিক্ষার্থীর বিদ্যালয় রোলঃ
বিদ্যালয়ের গবেষণা আইডিঃ
শিক্ষার্থীর গবেষণা আইডিঃ

(নিচের প্রশ্নগুলোতে যেখানে প্রযোজ্য সঠিক উত্তরে টিক চিহ্ন দাও। কিছু প্রশ্নের নিচে দেয়া খালি জায়গায় উত্তর লিখতে হবে। তুমি চাইলে হিসেব করার জন্য আলাদা কাগজ ব্যবহার করতে পার। এখানে একটি উদাহরণের সমাধান করা আছে।)

উদাহরণঃ
আকাশের ৫ টি খেলনা আছে। রিয়ার ৩ টি খেলনা আছে। তাদের একত্রে কয়টি খেলনা আছে?
ক) ৫
খ) ৩
✓ গ) ৮
ঘ) ২
ঙ) ৪

Initial Notes: Grade _____ Roll _____, Male/Female.

০১) জাহিদের কাছে ৫ টা আপেল ছিল। কলি তাকে আরো ৭ টা আপেল দিল। জাহিদের কাছে কতটুকু আপেল ছিল?
(ক) ১২ (খ) ৭ (গ) ৫ (ঘ) ২ (ঙ) উত্তর দেবার মত যথেষ্ট তথ্য নেই

Item 1 Notes:

০২) রাজির কাছে ১৫ টি মোমবাতি আছে। দাঁড়দের কাছে রাজির চেয়ে ৭ টি মোমবাতি কম আছে। দাঁড়দের কাছে কতগুলো মোমবাতি আছে সেটা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $১৫ + ৭$
(খ) $৭ + ১৫$
(গ) $১৫ - ৭$
(ঘ) $৭ - ১৫$
(ঙ) উপরের কোনটিই নয়

Item 2 Notes:

০৩) নিচের খালি জায়গায় এই যোগ অঙ্কটি করঃ ২৪ + ৩১।

Item 3 Notes:

০৪) $২২ + ৭ = ?$

- (ক) ২৭
(খ) ২৪
(গ) ২৯
(ঘ) ৩০
(ঙ) উপরের কোনোটিই নয়

Item 4 Notes:

০৫) মনে কর তোমাকে বলা হল ২৩৫ আর ৫৪ যোগ করতে। যোগ করে তুমি উত্তর পেলে ২৮৯। এই যোগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

- (ক)
$$\begin{array}{r} ২৩৫ \\ ৫৪+ \\ \hline ২৮৯ \end{array}$$

(খ)
$$\begin{array}{r} ২৮৯ \\ +৫৪ \\ \hline ২৩৫ \end{array}$$

(গ)
$$\begin{array}{r} ৫৪+ \\ ২৩৫ \\ \hline ২৮৯ \end{array}$$

(ঘ)
$$\begin{array}{r} ২৩৫ \\ +৫৪ \\ \hline ২৮৯ \end{array}$$

(ঙ) উপরের কোনোটিই নয়

Item 5 Notes:

০৬) ত্রিকোট খেলায় সুজন ১২০ রান, সুমন ৮৫ রান এবং সুজয় ৬৭ রান করল। তিনজনে মোট কত রান করল?

- (ক) ২০৫
- (খ) ২৭২
- (গ) ১৫২
- (ঘ) ৩৭২
- (ঙ) উপরের কোনটিই নয়

Item 6 Notes:

০৭) রাজুর তিনজন বন্ধু তাকে কিছু পয়সা দিল। প্রথমে সামিয়া তাকে ৪৫ টি পয়সা দিল, আকারিয়া তাকে ৮০ টি পয়সা দিল এবং তারপর লিসা তাকে ২৫ টি পয়সা দিল। বন্ধুদের মধ্যে শেষজন রাজুকে কয়টি পয়সা দিয়েছিল?

- (ক) ১২৫
- (খ) ৮০
- (গ) ২৫
- (ঘ) ৪৫
- (ঙ) কোনটিই নয়

Item 7 Notes:

০৮) সত্যিফের বাসায় ১৩৮ টি ফুলগাছ আছে। মিতুর বাসায় সত্যিফের চেয়ে ১১২ টি ফুলগাছ বেশী আছে। মিতুর বাসায় কতগুলো ফুলগাছ আছে সেটা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) ১৩৮ - ১২
- (খ) ১১২ - ১৩৮
- (গ) ১৪০ - ১২
- (ঘ) ১৪০ - ১৩৮
- (ঙ) উপরের কোনটিই নয়

Item 8 Notes:

৩৯) নিচের খালি জায়গায় এই বিয়োগ অঙ্কটি করয় ২৬ - ১২।

Item 9 Notes:

১০) $২৩ - ৮ = ?$

- (ক) ১৫
- (খ) ৩১
- (গ) ১৭
- (ঘ) ১৬
- (ঙ) উপরের কোনটিই নয়

Item 10 Notes:

১১) মনে কর, তোমাকে বলা হল ৪৩৯ থেকে ৮৪ বিয়োগ করতে। বিয়োগ করে উত্তর পাওয়া গেল ৩৫৫। এই বিয়োগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

- (ক)
$$\begin{array}{r} ৮৪ \\ - ৪৩৯ \\ \hline ৩৫৫ \end{array}$$
- (খ)
$$\begin{array}{r} ৪৩৯ \\ - ৮৪ \\ \hline ৩৫৫ \end{array}$$
- (গ)
$$\begin{array}{r} ৪৩৯ \\ - ৮৪ \\ \hline ৩৫৫ \end{array}$$

(খ) ৩৫৫

+৮৮

৪৩৯

(ঙ) উপরের কোনটিই

Item 11 Notes:

১২) অজয় বাবু ৪২৫০ টাকা নিয়ে বাজারে গেলেন। বাজার করার পর তার কাছে ৮৯০ টাকা রইল। তিনি কত টাকার বাজার করলেন?

(ক) ৫১৪০

(খ) ৮৯০

(গ) ৩৩৬০

(ঘ) ৪২১০

(ঙ) উপরের কোনটিই নয়

Item 12 Notes:

১৩) জামালের কাছে ৫ টি খেলনা আছে, বকরের কাছে ৮ টি খেলনা আছে। নিচের কোনটি সত্য?

(ক) জামালের বকরের চেয়ে বেশী খেলনা আছে

(খ) বকরের জামালের চেয়ে বেশী খেলনা আছে

(গ) বকরের জামালের চেয়ে কম খেলনা আছে

(ঘ) জামালা আর বকরের সমান সংখ্যক খেলনা আছে

(ঙ) উপরের কোনটিই নয়

Item 13 Notes:

১৪) স্কুল থেকে দুইশ তেইশজন ছাত্র, একশ পঁয়ত্রিশ জন ছাত্রী এবং বারজন শিক্ষক শিক্ষাসফরে যাচ্ছেন। মোট কতজন শিক্ষাসফরে যাচ্ছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

(ক) $২২৩ + ১৩৫$

(খ) $১৩৫ + ১২ - ২২৩$

(গ) $১২ + ১৩৫ + ২২৩$

(ঘ) $২২৩ - ১৩৫ - ১২$

(ঙ) উপরের কোনটিই

Item 14 Notes:

১৫) নিচের খালি জায়গায় এই যোগ্য অঙ্কটি করঃ $৫৮ + ২৭$ ।

Item 15 Notes:

১৬) নিচের কোন বিকল্পটি ৯৪ + ৩৩ এর যোগ্যফলের সবচেয়ে কাছাকাছি?

(ক) ১১০

(খ) ১৩০

(গ) ১২০

(ঘ) ১৪০

(ঙ) উপরের কোনটিই নয়

Item 16 Notes:

১৭) মনে কর তোমাকে বলা হল ২৬, ২৫৩৩ আর ৩২৫ যোগ করতে। যোগ করে তুমি উত্তর পেলে ২৮৮৪। এই যোগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

$$\begin{array}{r} (ক) \ ২৪৩৩ \\ \quad ২৬ \\ \hline + ৩২৫ \\ \hline ২৮৮৪ \end{array}$$

$$\begin{array}{r} (খ) \ ২৫৩৩ \\ \quad ৩২৫ \\ \hline - ২৬ \\ \hline ২৮৮৪ \end{array}$$

$$\begin{array}{r} (গ) \ ২৫৩৩ \\ \quad ৩২৫ \\ \hline + ২৬ \\ \hline ২৮৮৪ \end{array}$$

$$\begin{array}{r} (ঘ) \ ২৫৩৩ \\ \quad ৩২৫ + \\ \hline ২৬ \\ \hline ২৮৮৪ \end{array}$$

(ঙ) উপরের কোনটিই নয়

Item 17 Notes:

১৮) একটি শহরে ২৩৪৬৫ জন পুরুষ, ২৩৩৫ জন মহিলা ও ১১২৮৯ জন শিশু বাস করে। উক্ত শহরের জনসংখ্যা কত?

(ক) ৩৭০৮০

(খ) ৪০৮৩০

(গ) ৩৫৭৮০

(ঘ) ৩৬৯০০

(ঙ) উপরের কোনটিই নয়

Item 18 Notes:

১৯) একটি ত্রিকটি ক্রিকেট ব্যাটের ওজন ২ কিলোগ্রাম, একটি পেসিসের ওজন ৫০ গ্রাম, একটি রুনারের ওজন ১০০ গ্রাম, একটি টেলিভিশনের ওজন ৬ কিলোগ্রাম, এবং একটি টমটোর ওজন ৩০ গ্রাম। এগুলোর মধ্যে ক্রিকেট ব্যাটটি লম্বায় সবচেয়ে বড় কিন্তু কোন বস্তুর ওজন ক্রিকেট ব্যাটের চেয়ে বেশী?

- (ক) পেসিস
- (খ) রুনার
- (গ) টমটো
- (ঘ) টেলিভিশন
- (ঙ) উপরের কোনটিই নয়

Item 19 Notes:

২০) জবার কাছে কিছু কমলালেবু আছে। দিপুর কাছে ২৮ টি কমলালেবু আছে যা জবার চেয়ে ১২ টি বেশী। জবার কাছে ক'টি কমলালেবু আছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $২৮ + ১২$
- (খ) $১২ + ২৮$
- (গ) $২৮ - ১২$
- (ঘ) $৪০ - ১২$
- (ঙ) উপরের কোনটিই নয়

Item 20 Notes:

২১) নিচের খালি জায়গায় এই বিয়োগ অঙ্কটি করঃ ৬৩ - ১৯ ।

Item 21 Notes:

২২) নিচের কোন বিকল্পটি ৮৭ - ৫৩ এর বিয়োগফলের সবচেয়ে কাছাকাছি?

- (ক) ৩৭
- (খ) ৪৫
- (গ) ২৭
- (ঘ) ৩০
- (ঙ) উপরের কোনটিই নয়

Item 22 Notes:

২৩) মনে কর, তোমাকে বলা হল ২৬১৩ থেকে ১২৯ বিয়োগ করতে। বিয়োগ করে উত্তর পাওয়া গেল ২৪৮৪। এই বিয়োগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

- (ক)
$$\begin{array}{r} ২৬৩১ \\ - ১২৯ \\ \hline ২৪৮৪ \end{array}$$
- (খ)
$$\begin{array}{r} ২৬১৩ \\ - ১২৯ \\ \hline ২৪৮৪ \end{array}$$
- (গ)
$$\begin{array}{r} ২৬১৩ \\ - ১২৯ \\ \hline ২৪৮৪ \end{array}$$
- (ঘ)
$$\begin{array}{r} ২৬১৩ \\ - ১২৯ \\ \hline ২৪৮৪ \end{array}$$
- (ঙ) উপরের কোনটিই নয়

Item 23 Notes:

২৪) বার্ষিক ক্রীড়া প্রতিযোগিতার জন্য ৯৫০০ টাকা ধরোজন। সরকারি অনুদান হিসেবে ৪৫০০ টাকা এবং বিদ্যালয়ের তহবিল থেকে ২০০০ টাকা দেয়া হয়েছে। অনুষ্ঠানটি করার জন্য আর কত টাকা ঘোষণা করতে হবে?

- (ক) ৪৫০০
- (খ) ২০০০
- (গ) ৩০০০
- (ঘ) ৫০০০
- (ঙ) উপরের কোনটিই নয়

Item 24 Notes:

২৫) খুলনা ঢাকা থেকে পশ্চিম দিকে অবস্থিত, নরসিংদী ঢাকা থেকে পূর্ব দিকে অবস্থিত। সিলেট ঢাকা থেকে আরও পূর্ব দিকে অবস্থিত। এই শহরগুলোর মধ্যে কোনটি সর্বাধিক পশ্চিমে অবস্থিত?

- (ক) সিলেট
- (খ) ঢাকা
- (গ) নরসিংদী
- (ঘ) খুলনা
- (ঙ) উপর দেবার মত যথেষ্ট তথ্য নেই

Item 25 Notes:

২৬) রাহাতের কাছে ৪১ টি ডাকটিকিট আছে, রুবিনের কাছে ৩০ টি ডাকটিকিট আছে। লিসার কাছে রাহাত আর রুবিন দু'জনের মিলে যতগুলো ডাকটিকিট তার চেয়ে ২৩ টি ডাকটিকিট বেশী আছে। লিসার কাছে কতটি ডাকটিকিট আছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $৪১ + ৩০ - ২৩$
- (খ) $৪১ + ৩০ + ২৩$
- (গ) $৩০ - ২৩ + ৪১$
- (ঘ) $৪১ - ৩০ - ২৩$
- (ঙ) উপরের কোনটিই নয়

Item 26 Notes:

২৭) নিচের খালি জায়গায় এই যোগ অঙ্কটি করঃ $১০৩৯ + ১৪২ + ১৪$ ।

Item 27 Notes:

২৮) নিচের কোন বিকল্পটি $২১৩৭ + ২৫১$ এর যোগফলের সবচেয়ে কাছাকাছি?

- (ক) ২৩৫০
- (খ) ২৩৬০
- (গ) ২৪০০
- (ঘ) ২৪২০
- (ঙ) উপরের কোনটিই নয়

Item 28 Notes:

২৯) একজন শিক্ষক তার স্কুলের জন্য নতুন আসবাবপত্র কিনেছেন। তিনি দু'টো টেবিল কিনেছেন যেগুলোর প্রতিটির দৈর্ঘ্য ৭ ফুট এবং ওজন ২৫ কিলোগ্রাম। এছাড়াও তিনি ৫ টি বেঞ্চ কিনেছেন যেগুলোর প্রতিটির দৈর্ঘ্য ৫ ফুট এবং ওজন ২০ কিলোগ্রাম। শিক্ষক হিসেব করে দেখলেন আসবাবপত্রগুলোর মোট ওজন ১৫০ কিলোগ্রাম এবং সেগুলোকে স্কুল নিয়ে যাবার জন্য একটি বাহনের ব্যবস্থা করলেন। নিচের বিকল্পগুলোর মধ্যে কোনটিতে সবচেয়ে সঠিক হিসেব এবং উপযুক্ত পরিবহন মাধ্যম রয়েছে?

- (ক) ওজন $২৫+ ২৫+ ২৫+ ২৫+ ২০+ ২০+ ২০= ১৫০$ কিলোগ্রাম; পরিবহনের জন্য রিকশা
 (খ) ওজন $২৫+ ২৫+ ২৫+ ২৫+ ২০+ ২০+ ২০+ ২০= ১৫০$ কিলোগ্রাম; পরিবহনের জন্য রিকশা ভ্যান
 (গ) ওজন $২৫+ ২৫+ ২০+ ২০+ ২০+ ২০+ ২০= ১৫০$ কিলোগ্রাম; পরিবহনের জন্য ট্রাক
 (ঘ) ওজন $২৫+ ২৫+ ২০+ ২০+ ২০+ ২০+ ২০= ১৫০$ কিলোগ্রাম; পরিবহনের জন্য বাস
 (ঙ) উপরের কোনটিই

Item 29 Notes:

৩০) কোন শহরে নারী পুরুষের সংখ্যা ৪২৫৩০, এদের মধ্যে ২১৬৭৮ জন পুরুষ হলে, নারীর সংখ্যা নির্ণয় কর।

- (ক) ১১৫৬১
 (খ) ২০৮৫২
 (গ) ২০৫২১
 (ঘ) ১১৩২২
 (ঙ) উপরের কোনটিই নয়

Item 30 Notes:

৩১) সুইমিং পুলের পানিতে ঝানিকটা লবণ থাকে। কলের পানিতে সবচেয়ে কম লবণ থাকে। তবে আটলান্টিক মহাসাগর বা প্রশান্ত মহাসাগরের পানিতে ধূসর লবণ থাকে। কোন পানিতে সবচেয়ে বেশী লবণ থাকে?

- (ক) কলের পানিতে
 (খ) সুইমিং পুলের পানিতে
 (গ) আটলান্টিক মহাসাগরের পানিতে
 (ঘ) প্রশান্ত মহাসাগরের পানিতে

(ঙ) উত্তর দেবার মত যথেষ্ট তথ্য নেই

Item 31 Notes:

৩২) নতুন ক্লাসের বই-খাতা কেনার জন্য রুবি বাসা থেকে ২৫০০ টাকা পেয়েছে। নুসরাত রুবির চেয়ে ৭০০ টাকা কম পেয়েছে আর রিমা নুসরাতের চেয়ে ৩০০ টাকা কম পেয়েছে। রিমা কত টাকা পেয়েছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $২৫০০ + ৭০০ + ৩০০$
(খ) $২৫০০ - ৭০০ + ৩০০$
(গ) $২৫০০ - ৭০০ - ৩০০$
(ঘ) $৭০০ + ৩০০ - ২৫০০$
(ঙ) উপরের কোনটিই নয়

Item 32 Notes:

৩৩) নিচের বাগি জায়গায় এই বিয়োগ অঙ্কটি করঃ ১২৬৯ - ২৩৩ - ১২।

Item 33 Notes:

৩৪) ১২৪৫ - ১১৮ এর বিয়োগফলের জন্য সঠিক উত্তর নিচের কোনটি?

- (ক) ১১২৭
(খ) ১১১৭
(গ) ১১৩৭
(ঘ) ১০২৩
(ঙ) উপরের কোনটিই নয়

Item 34 Notes:

৩৫) একটি একটি স্কুলের প্রধান শিক্ষক স্কুলে বিজয় দিবস উদযাপনের আয়োজন করতে চাইছেন যেখানে সব ছাত্র-ছাত্রীরা হাতে বাংলাদেশের একটি করে ছোট পতাকা থাকবে। এই পতাকাগুলো ২০০ টির প্যাকেটে পাওয়া যায়। স্কুলে এমন ৪ টি প্যাকেট আছে। ওই স্কুলে মোট ৬৫৬ জন ছাত্র-ছাত্রী আছে কিন্তু তাদের মধ্যে ১২৬ জন বিজয় দিবসের অনুষ্ঠানে উপস্থিত থাকতে পারবে না। প্রধান শিক্ষক হিসেব করে দেখলেন অনুষ্ঠানে মোট ৫৩০ জন ছাত্র-ছাত্রী থাকবে এবং নতুন করে পতাকাও কিনতে হবে না। নিচের কোন বিকল্পটিতে অনুষ্ঠানে উপস্থিত ছাত্র-ছাত্রী সংখ্যা ও প্রয়োজনীয় পতাকার সংখ্যা বের করার জন্য সঠিক গণনা এবং উত্তর আছে?

- (ক) $৬৫৬ + ১২৬ = ৫৩০$; ৪ প্যাকেট পতাকা
(খ) $৬৫৬ - ১২৬ = ৫৩০$; ৪ প্যাকেট পতাকা
(গ) $৬৫৬ - ১২৬ = ৫৩০$; ৩ প্যাকেট পতাকা
(ঘ) $৬৫৬ - ১২৬ = ৫৩০$; ২ প্যাকেট পতাকা
(ঙ) উপরের কোনটিই নয়

Item 35 Notes:

৩৬) সিয়াম অপেক্ষা শিহাবের ৩৯০ টাকা বেশি এবং শিমুল অপেক্ষা ৪৭০ টাকা কম আছে। শিমুলের ৮৯০ টাকা আছে। সিয়াম ও শিহাবের প্রত্যেকের কত টাকা আছে?

- (ক) সিয়াম ৪৭০ টাকা, শিহাব ৩৯০ টাকা
(খ) সিয়াম ৮১০ টাকা, শিহাব ৪২০ টাকা
(গ) সিয়াম ৪২০ টাকা, শিহাব ৮১০ টাকা
(ঘ) সিয়াম ৮৬০ টাকা, শিহাব ৮৯০ টাকা
(ঙ) উপরের কোনটিই নয়

Item 36 Notes:

৩৭) পিয়াকে পিয়াকে নিচের অঙ্কটি করতে দেয়া হয়েছিল:

'হাশেদ স্কুলে যাবার জন্য বাসা থেকে সকাল ১০.২০ এ বের হয়ে সকাল ১১.৩০ এ ক্লাসে পৌঁছাল। তার ক্লাসে পৌঁছতে কত সময় লেগেছে?'

পিয়া অঙ্কটা করে উত্তর পেল ৫০। উত্তরটা সে কিতাবে লিখবে?

- (ক) সকাল ৫০
- (খ) ৫০ ঘন্টা
- (গ) ৫০ মিনিট
- (ঘ) ০.৫০ ঘন্টা
- (ঙ) উপরের কোনটিই নয়

Item 37 Notes:

৩৮) লিয়নকে নিচের অঙ্কটি করতে দেয়া হয়েছিল:

'জেনি ১২.৫ টাকা দিয়ে ৫ টি চকলেট কিনেছে। সে দোকানীকে ২০ টাকার একটা নোট দিল। দোকানী তাকে কত ফেরত দেবে?'

লিয়ন অঙ্কটা করে উত্তর পেল ৭.৫। উত্তরটা সে কিতাবে লিখবে?

- (ক) ৭ টাকা
- (খ) ৭ টাকা ৫০ পয়সা
- (গ) ৭.৫ পয়সা
- (ঘ) ৭.৫
- (ঙ) উপরের কোনটিই নয়

Item 38 Notes:

৩৯) ২৫ + ৩২ এই যোগ অঙ্কটি তুমি কিতাবে করবে চিন্তা কর। নিচের কোন বিকল্পটি তোমার চিন্তার সাথে সবচেয়ে বেশী মেলে?

- (ক) তুমি উপর-নিচে যোগ করবে
- ২৫
- + ৩২
- ৫৭

- (খ) $২৫ + ৩২ = ৫৭$ হয় এটা তুমি মুখস্থ জানো
 (গ) তুমি মনে মনে হিসেব করবে যে $২০ + ৩০ = ৫০$ আর $৫ + ২ = ৭$; এর পরে $৫০ + ৭ = ৫৭$
 (ঘ) তুমি মনে মনে হিসেব করবে যে $৩২ + ৮ = ৪০$ আর $২৫ - ৮ = ১৭$; এর পরে $৪০ + ১৭ = ৫৭$
 (ঙ) এই অঙ্কটি বিভাবে করতে হবে তুমি নিশ্চিত নও
 (চ) তুমি অন্য কোন পদ্ধতি ব্যবহার করবে। কী পদ্ধতি তা নিচের খালি জায়গায় লিখ:

Item 39 Notes:

৪০) ১৪ - ৮ এই বিয়োগ অঙ্কটি তুমি কভাবে করবে চিন্তা কর। নিচের কোন বিকল্পটি তোমার চিন্তার সাথে সবচেয়ে বেশী মেলে?

- (ক) তুমি আন্দাজে উল্লর দেবে/ অঙ্কটি করার পদ্ধতি তোমার জানা নেই
 (খ) $১৪ - ৮ = ৬$ হয় এটা তুমি মুখস্থ জানো
 (গ) তুমি ৮ থেকে ৮, ৯, ১০ এভাবে ১৪ পর্যন্ত গুনাতে থাকবে এবং উল্লর ৬ হয় সেটা বের করবে
 (ঘ) তুমি জানো $৮ = ৪ + ৪$ আর $১৪ - ৪ = ১০$, এরপর $১০ - ৪ = ৬$
 (ঙ) তুমি উপর-নিচে বিয়োগ করবে

$$\begin{array}{r} ১৪ \\ - ৮ \\ \hline ৬ \end{array}$$

- (চ) তুমি অন্য কোন পদ্ধতি ব্যবহার করবে। কী পদ্ধতি তা নিচের খালি জায়গায় লিখ:

Item 40 Notes:

Additional Questions:

১. এই পরীক্ষার কোন প্রশ্নের উত্তর দিতে কি তুমি সমস্যায় পড়েছ? কোন প্রশ্ন/প্রশ্নগুলো? কি ধরনের সমস্যায় পড়েছ?

Additional Q1 Notes:

২. যোগ বিয়োগ করতে গিয়ে তুমি কী ধরনের কৌশল ব্যবহার কর? এদের মধ্যে কোন কৌশল কি বেশী ব্যবহার কর?

Additional Q2 Notes:

৩. কিসের উপর ভিত্তি করে তুমি কৌশল নির্ধারণ কর?

Additional Q3 Notes:

৪. তুমি ২৫+৩৭ এই যোগটি কিভাবে করবে? ধাপে ধাপে আমাকে বল।

Additional Q4 Notes:

৫. তুমি ৫৪-২৮ এই বিয়োগটি কিভাবে করবে? ধাপে ধাপে আমাকে বল।

Additional Q5 Notes:

৬. কোন একটি গাণিতিক সমস্যার যে উত্তর তুমি পেয়েছ সেটি সঠিক কী না তা তুমি কিভাবে যাচাই কর?

Additional Q6 Notes:

৭. পাটিগণিতের শব্দ-সম্মিলিত সমস্যা বা বাক্যক সমাধান করার সময় তোমার কোন ভুল সবচেয়ে বেশী হয় বলে তুমি মনে কর?

Additional Q7 Notes:

Appendix 3.23: Students' cognitive interview fieldnote template for the final study (English)

Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

Student school S/N:
School research ID:
Student research ID:

Example: Akash has 5 toys. Ria has 3 toys. How many toys do they have together?

- a. 5
- b. 3
- ✓ c. 8
- d. 2
- e. 4
- f. none of the above

Initial notes: Grade _____, Class Roll _____, Male/Female

(The interviewer explains how the think-aloud is going to take place. He uses the example item above to demonstrate how to articulate thinking while answering an item. Then the respondent is requested to start answering the test items. Respondents are given blank papers in case they want to make notes or small calculations. There is space under each item for the interviewer to record the responses. Additional paper can also be used for note taking.)

1. Jahid had 5 apples. Koli gave him 7 more apples. How many apples did Jahid have in the beginning?

- a. 12 b. 7 c. 5 d. 2 e. Not enough information to tell

Item 1 Notes:

2. Razia has 15 candles. Daud has 7 candles less than Razia. Which of these number sentences is appropriate for calculating how many candles Daud has?

- a. $15 + 7$
- b. $7 + 15$
- c. $7 - 15$
- d. $15 - 7$
- e. None of the above

Item 2 Notes:

3. Solve this addition problem $2457 + 1319$. Show your work.

Item 3 Notes:

4. $522 + 87 = ?$

- a. 527
- b. 609
- c. 435
- d. 509
- e. None of the above

Item 4 Notes:

5. Suppose you were asked to add 235 and 54, and you found the result was 289. How would you write the solution down?

a. 235

$$\begin{array}{r} 54+ \\ \hline 289 \end{array}$$

b. 289

$$\begin{array}{r} +54 \\ \hline 235 \end{array}$$

c. 54+

$$\begin{array}{r} 234 \\ \hline 289 \end{array}$$

d. 235

$$\begin{array}{r} +54 \\ \hline 289 \end{array}$$

e. None of the above

Item 5 Notes:

6. *(Item removed from cognitive interview as it is a textbook item)*

7. **3 of Raju's friends gave him some coins. First, Samia gave him 45 coins; Jakaria gave him 80 coins; and then Lisa gave him 25 coins. How many coins did he get from the last one of his friends?**

a. 125

b. 80

c. 45

d. 25

e. None of the above

Item 7 Notes:

8. Latif has 138 plants in his house. Mitu has 112 plants more than him. Which of these number sentences would be appropriate for calculating how many plants Mitu has?

- a. $138 - 12$
- b. $112 - 138$
- c. $140 - 12$
- d. $140 - 138$
- e. None of the above

Item 8 Notes:

9. Solve this subtraction problem $4869 - 1344$. Show your work.

Item 9 Notes:

10. $328 - 75 = ?$

- a. 253
- b. 403
- c. 263
- d. 273
- e. None of the above

Item 10 Notes:

11. Suppose you were asked to subtract 84 from 439 and you found the answer was 355. How would you write the solution down?

a.
$$\begin{array}{r} 84 \\ - 439 \\ \hline 355 \end{array}$$

b.
$$\begin{array}{r} 439 \\ - 84 \\ \hline 355 \end{array}$$

c.
$$\begin{array}{r} 439 \\ 84 - \\ \hline 355 \end{array}$$

d.
$$\begin{array}{r} 355 \\ + 84 \\ \hline 439 \end{array}$$

e. None of the above

Item 11 Notes:

12. *(Item removed from cognitive interview as it is a textbook item)*

13. Jamal has more toys than Bakar who has more toys than Rimi but less than Rafi. Which of the following is true?

- a. Bakar has more toys than Rafi
- b. Rafi has more toys than Rimi
- c. Rimi has more toys than Rafi
- d. Rafi has more toys than Bakar
- e. None of the above

Item 13 Notes:

14. Two hundred and twenty-three boys, one hundred and thirty-five girls, and twelve teachers participated in a school trip. Which of these number sentences is correct for finding out the total number of people going on the trip?

- a. $223 + 135$
- b. $135 + 12 - 223$
- c. $223 - 135 - 12$
- d. $12 + 135 + 223$
- e. None of the above

Item 14 Notes:

15. Solve this addition problem $52758 + 25936$. Show your work.

Item 15 Notes:

16. Which of the following is the closest to the result for solving $94 + 33$?

- a. 110
- b. 130
- c. 120
- d. 140
- e. None of the above

Item 16 Notes:

17. Suppose you were asked to add 26, 2533, and 325. You found the answer was 2884. How would you write the solution down?

a.
$$\begin{array}{r} 2433 \\ 26 \\ +325 \\ \hline 2884 \end{array}$$

b.
$$\begin{array}{r} 2533 \\ 325 \\ - 26 \\ \hline 2884 \end{array}$$

c.
$$\begin{array}{r} 2533 \\ 325 \\ + 26 \\ \hline 2884 \end{array}$$

d.
$$\begin{array}{r} 2533 \\ 325+ \\ 26 \cdot \\ \hline 2884 \end{array}$$

e. None of the above

Item 17 Notes:

18. *(Item removed from cognitive interview as it is a textbook Item)*

19. **A cricket bat weighs 2 kilograms, a pencil weighs 50 grams, a ruler weighs 100 grams, a television weighs 6 kilograms, and a tomato weighs 30 grams. A cricket bat has the highest length among these but which of these objects weighs more than the cricket bat?**

- a. Pencil
- b. Ruler
- c. Tomato
- d. Television

Item 19 Notes:

- e. None of the above
20. **Joba has some oranges. Dilu has 28 oranges which is 12 more than Joba. Which of the following number sentences is right for figuring out how many oranges Joba has?**
- a. $28 + 12$
 - b. $12 - 28$
 - c. $12 + 28$
 - d. $28 - 12$
 - e. None of the above

Item 20 Notes:

21. **Solve this subtraction problem $6867 - 2593$. Show your work.**

Item 21 Notes:

22. **Which of the following is the closest to the result for solving $87 - 53$?**
- a. 37
 - b. 44
 - c. 27
 - d. 30
 - e. None of the above

Item 22 Notes:

23. Suppose you were asked to subtract 129 from 2613 and you found the answer was 2484. How would you write the solution down?

a.
$$\begin{array}{r} 2631 \\ -129 \\ \hline 2484 \end{array}$$

b.
$$\begin{array}{r} -129 \\ 2613 \\ \hline 2884 \end{array}$$

c.
$$\begin{array}{r} 2613 \\ -129 \\ \hline 2484 \end{array}$$

d.
$$\begin{array}{r} 2613 \\ 129- \\ \hline 2484 \end{array}$$

e. None of the above

Item 23 Notes:

24. *(Item removed from cognitive interview as it is a textbook item)*

25. Khulna is to the west from Dhaka, Narsingdi is to the east of Dhaka. Sylhet is further east from Dhaka. Which of these cities is the furthest towards the west?

- a. Sylhet
- b. Dhaka
- c. Narsingdi
- d. Khulna
- e. Not enough information to tell

Item 25 Notes:

26. Rahat has 41 stamps, Rubin has 30 stamps. Lisa has 23 stamps more than what Rahat and Rubin have together. Which of the following number sentences is right for calculating how many stamps Lisa has?

- a. $41 + 30 - 23$
- b. $41 + 30 + 23$
- c. $30 - 23 + 41$
- d. $41 - 30 - 23$
- e. None of the above

Item 26 Notes:

27. Solve this addition problem $40695 + 17582 + 10346$. Show your work.

Item 27 Notes:

28. Which of the following is the closest to the result for solving for $2137 + 251$?

- a. 2350
- b. 2360
- c. 2420
- d. 2400
- e. None of the above

Item 28 Notes:

29. A teacher is buying some new furniture for his school. He bought two tables which are 7 feet long and weighs 25 kilograms each. He also bought 5 desks, each of which are 5 feet long and weighs 20 kilograms. The teacher

calculated the total weight of the furniture to be 150 kg and decided a method of transportation. Which of the following alternatives show the correct calculation and an appropriate transportation method?

- a. Weight $25+25+25+25+20+20+20= 150$ kilogram; a rickshaw for transportation
- b. Weight $25+25+25+25+20+20+20= 150$ kilogram; a truck for transportation
- c. Weight $25+25+20+20+20+20+20 = 150$ kilogram; a truck for transportation
- d. Weight $25+25+20+20+20+20+20 = 150$ kilogram; a bus for transportation
- e. None of the above

Item 29 Notes:

30. *(Item removed from cognitive interview as it is a textbook Item)*

31. Water from a swimming pool has some salt in it. Tap water has the least amount of salt. But there is much salt in water from the Atlantic or the Pacific Ocean. Which water contains the most salt?

- a. Tap water
- b. Swimming pool water
- c. Water from the Atlantic
- d. Water from the Pacific
- e. Not enough information to tell

Item 31 Notes:

32. Rubi got 2500 taka from her parents to buy new books. Nusrat got 700 taka less than Rubi and Rima got 300 taka less than Nusrat. Which of the following number sentences is correct for calculating the amount Rima got?

- a. $2500 + 700 + 300$
- b. $2500 - 700 + 300$
- c. $2500 + 700 - 300$

- d. $2500 - 700 - 300$
- e. None of the above

Item 32 Notes:

33. Solve this subtraction problem $4912 - 2385 - 1456$. Show your work.

Item 33 Notes:

34. Which of the following is the correct answer for solving $7322 - 3980$?

- a. 3342
- b. 4342
- c. 3242
- d. 4242
- e. None of the above

Item 34 Notes:

35. The headteacher of a school is planning a victory day celebration in her school and wanted each student to have a small Bangladeshi flag. These flags come in packets of 200s. There are 4 packets of flags in the school. The school has 656 students in total but 126 of them will not be present during the celebration. The headteacher calculated that there is going to be 530 students in the celebration and the school would not need to buy new flags.

Which of the following alternatives show the correct calculation and result for the number of students and number of packets of flags required?

- a. $656 + 126 = 530$; 4 packets of flag
- b. $656 - 126 = 530$; 4 packets of flag
- c. $656 - 126 = 530$; 3 packets of flag
- d. $656 - 126 = 530$; 2 packets of flag
- e. None of the above

Item 35 Notes:

36. *(Item removed from cognitive interview as it is a textbook item)*

37. Pia was asked to solve this problem: 'Rashed started for his school from home at 10.40 am and reached the class at 11.30 am. How long did it take for him to reach his class?'

She solved the problem and found the answer was 50. How shall she write down the result?

- a. 50 am
- b. 50 hours
- c. 50 minutes
- d. 0.50 hours
- e. None of the above

Item 37 Notes:

38. Leon was asked to solve the following problem: 'Jenny bought 5 candies for 12.5 taka. She gave the shopkeeper a 20-taka note. How much shall she get in change?'

He solved the problem and found the answer was 7.5. How should he write down the answer?

- a. 7 taka
- b. 7 taka and 50 poisa
- c. 7.5 poisa
- d. 7.5
- e. None of the above

Item 38 Notes:

Additional Questions:

1. Did you find any item of this test problematic while answering? Which item(s)?
What problem did you face?

Additional Q1 Notes:

2. What strategies do you use while doing addition and subtraction? Are there certain strategies that you use more?

Additional Q2 Notes:

3. How do you decide which strategy/strategies to use?

Additional Q3 Notes:

4. How would you do (this addition) $25+37$? Tell me step-by-step.

Additional Q4 Notes:

5. How would you do (this subtraction) $54-28$? Tell me step-by-step.

Additional Q5 Notes:

6. After solving a mathematical problem, how do you check whether the answer you have got is correct?

Additional Q6 Notes:

7. What mistakes do you think you make the most while solving arithmetic word problems?

Additional Q7 Notes:

Appendix 3.24: Students' cognitive interview fieldnote template for the final data collection (Bangla)

Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

শিক্ষার্থীর বিদ্যালয় রোলস্ট
বিদ্যালয়ের গবেষণা আইডিঃ
শিক্ষার্থীর গবেষণা আইডিঃ

(নিচের প্রশ্নগুলোতে যেখানে প্রযোজ্য সঠিক উত্তরে টিক চিহ্ন দাও। কিছু প্রশ্নের নিচে দেয়া খালি জায়গায় উত্তর লিখতে হবে। তুমি চাইলে হিসেব করার জন্য আলাদা কাগজ ব্যবহার করতে পার। এখানে একটি উদাহরণের সমাধান করা আছে।)

উদাহরণঃ
আকাশের ৫ টি বেলনা আছে। রিয়ার ৩ টি বেলনা আছে। তাদের একত্রে কয়টি বেলনা আছে?
ক) ৫
খ) ৩
✓ গ) ৮
ঘ) ২
ঙ) ৪

Initial Notes: School: _____ Grade _____ Roll _____, Male/Female.

০১) জাহিদের কাছে ৫ টা আপেল ছিল। কলি তাকে আরো ৭ টা আপেল দিল। জাহিদের কাছে কতটুকু কটা আপেল ছিল?
(ক) ১২ (খ) ৭ (গ) ৫ (ঘ) ২ (ঙ) উত্তর দেবার মত যথেষ্ট তথ্য নেই

Item 1 notes:

০২) রাজিয়ার কাছে ১৫ টি মোমবাতি আছে। দাউদের কাছে রাজিয়ার চেয়ে ৭ টি মোমবাতি কম আছে। দাউদের কাছে কতগুলো মোমবাতি আছে সেটা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

(ক) $15 + 7$
(খ) $7 + 15$
(গ) $7 - 15$
(ঘ) $15 - 7$
(ঙ) উপরের কোনটিই নয়

Item 2 notes:

০৩) নিচের বালি জায়গায় এই যোগ অঙ্কটি করঃ $২৪৫৭ + ১৩১৯$ ।

Item 3 notes:

০৪) $৫২২ + ৮৭ = ?$

- (ক) ৫২৭
- (খ) ৬০৯
- (গ) ৪৩৫
- (ঘ) ৫০৯
- (ঙ) উপরের কোনটিই নয়

Item 4 notes:

০৫) মনে কর তোমাকে বলা হল ২৩৫ আর ৫৪ যোগ করতে। যোগ করে তুমি উত্তর পেলে ২৮৯। এই যোগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

- (ক)
$$\begin{array}{r} ২৩৫ \\ ৫৪+ \\ \hline ২৮৯ \end{array}$$
- (খ)
$$\begin{array}{r} ২৮৯ \\ +৫৪ \\ \hline ২৩৫ \end{array}$$
- (গ)
$$\begin{array}{r} ৫৪+ \\ ২৩৫ \\ \hline ২৮৯ \end{array}$$
- (ঘ)
$$\begin{array}{r} ২৩৫ \\ +৫৪ \\ \hline ২৮৯ \end{array}$$
- (ঙ) উপরের কোনটিই নয়

Item 5 notes:

০৭) রাজুর তিনজন বন্ধু তাকে কিছু পয়সা দিল। প্রথমে সামিয়া তাকে ৪৫ টি পয়সা দিল, জাকারিয়া তাকে ৮০ টি পয়সা দিল এবং তারপর লিসা তাকে ২৫ টি পয়সা দিল। বন্ধুদের মধ্যে শেষজন রাজুকে কয়টি পয়সা দিয়েছিল?

- (ক) ১২৫ (খ) ৮০ (গ) ৪৫ (ঘ) ২৫ (ঙ) কোনটিই নয়

Item 7 notes:

০৮) লতিফের বাসায় ১৩৮ টি ফুলগাছ আছে। মিতুর বাসায় লতিফের চেয়ে ১১২ টি ফুলগাছ বেশী আছে। মিতুর বাসায় কতগুলো ফুলগাছ আছে সেটা হিসেব করার জন্য নিচের কোন গাণিতিক বিকৃতিটি সঠিক হবে?

- (ক) ১৩৮ - ১২
(খ) ১১২ - ১৩৮
(গ) ১৪০ - ১২
(ঘ) ১৪০ - ১৩৮
(ঙ) উপরের কোনটিই নয়

Item 8 notes:

০৯) নিচের খালি জায়গায় এই বিরোধ অঙ্কটি কর ৪৮৬৯ - ১৩৪৪।

Item 9 notes:

১০) $৩২৮ - ৭৫ = ?$

- (ক) ২৫৩
(খ) ৪০৩
(গ) ২৬৩
(ঘ) ২৭৩
(ঙ) উপরের কোনটিই নয়

Item 10 notes:

১১) মনে কর, তোমাকে বলা হল ৪৩৯ থেকে ৮৪ বিয়োগ করতে। বিয়োগ করে উত্তর পাওয়া গেল ৩৫৫। এই বিয়োগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

- (ক) $\begin{array}{r} ৮৪ \\ - ৪৩৯ \\ \hline ৩৫৫ \end{array}$
(খ) $\begin{array}{r} ৪৩৯ \\ - ৮৪ \\ \hline ৩৫৫ \end{array}$
(গ) $\begin{array}{r} ৪৩৯ \\ ৮৪- \\ \hline ৩৫৫ \end{array}$
(ঘ) $\begin{array}{r} ৩৫৫ \\ + ৮৪ \\ \hline ৪৩৯ \end{array}$
(ঙ) উপরের কোনটিই নয়

Item 11 notes:

১২) জামালের বকরের চেয়ে বেশী খেলনা আছে। বকরের রিমির চেয়ে বেশী কিন্তু রাফির চেয়ে কম খেলনা আছে। নিচের কোনটি সত্য?

- (ক) বকরের রাফির চেয়ে বেশী খেলনা আছে
(খ) রাফির রিমির চেয়ে বেশী খেলনা আছে
(গ) রিমির রাফির চেয়ে বেশী খেলনা আছে

- (খ) রাফির বন্ধরের চেয়ে বেশী খেলনা আছে
(ঙ) উপরের কোনটিই নয়

Item 13 notes:

- ১৪) একটি স্কুল থেকে দুইশ তেইশজন ছাত্র, একশ পঁয়ত্রিশ জন ছাত্রী এবং বারজন শিক্ষক শিক্ষাসফরে যাচ্ছেন। মোট কতজন শিক্ষাসফরে যাচ্ছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?
- (ক) $২২৩ + ১৩৫$
(খ) $১৩৫ + ১২ - ২২৩$
(গ) $২২৩ - ১৩৫ - ১২$
(ঘ) $১২ + ১৩৫ + ২২৩$
(ঙ) উপরের কোনটিই নয়

Item 14 notes:

- ১৫) নিচের খালি জায়গায় এই যোগ্য অঙ্কটি করঃ $৫২৭৫৮ + ২৫৯৩৬$ ।

Item 15 notes:

১৬) নিচের কোন বিকল্পটি ৯৪ + ৩৩ এর যোগফলের সবচেয়ে কাছাকাছি?

- (ক) ১১০
(খ) ১৩০
(গ) ১২০
(ঘ) ১৪০
(ঙ) উপরের কোনটিই নয়

Item 16 notes:

১৭) মনে কর তোমাকে বলা হল ২৬, ২৫৩৩ আর ৬২৫ যোগ করতে। যোগ করে তুমি উত্তর পেলে ২৮৮৪। এই যোগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

- (ক)
$$\begin{array}{r} ২৬ \\ + ৬২৫ \\ \hline ২৮৮৪ \end{array}$$

(খ)
$$\begin{array}{r} ২৫৩৩ \\ ৬২৫ \\ \hline = ২৬ \\ ২৮৮৪ \end{array}$$

(গ)
$$\begin{array}{r} ২৫৩৩ \\ ৬২৫ \\ \hline + ২৬ \\ ২৮৮৪ \end{array}$$

(ঘ)
$$\begin{array}{r} ২৫৩৩ \\ ৬২৫ + \\ \hline ২৬ \\ ২৮৮৪ \end{array}$$

(ঙ) উপরের কোনটিই নয়

Item 17 notes:

১৯) একটি ক্রিকেট ব্যাটের ওজন ২ কিলোগ্রাম, একটি পেল্লের ওজন ৫০ গ্রাম, একটি রুলারের ওজন ১০০ গ্রাম, একটি টেলিভিশনের ওজন ৬ কিলোগ্রাম, এবং একটি টমিটোর ওজন ৩০ গ্রাম। এগুলোর মধ্যে ক্রিকেট ব্যাটটি লম্বায় সবচেয়ে বড় কিন্তু কোন বস্তুর ওজন ক্রিকেট ব্যাটের চেয়ে বেশী?

- (ক) পেল্লা
- (খ) রুলার
- (গ) টমিটো
- (ঘ) টেলিভিশন
- (ঙ) উপরের কোনটিই নয়

Item 19 notes:

২০) জবার কাছে কিছু কমলালেবু আছে। দিলুর কাছে ২৮ টি কমলালেবু আছে যা জবার চেয়ে ১২ টি বেশী। জবার কাছে ক'টি কমলালেবু আছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $২৮ + ১২$
- (খ) $১২ - ২৮$
- (গ) $১২ + ২৮$
- (ঘ) $২৮ - ১২$
- (ঙ) উপরের কোনটিই নয়

Item 20 notes:

২১) নিচের বালি জায়গায় এই বিয়োগ অঙ্কটি করঃ ৬৮৬৭ - ২৫৯৩ ।

Item 21 notes:

২২) নিচের কোন বিকল্পটি ৮৭ - ৫৩ এর বিয়োগফলের সবচেয়ে কাছাকাছি?

- (ক) ৩৭
- (খ) ৪৪
- (গ) ২৭
- (ঘ) ৩০
- (ঙ) উপরের কোনটিই নয়

Item 22 notes:

২৩) মনে কর, তোমাকে বলা হল ২৬১৩ থেকে ১২৯ বিয়োগ করতে। বিয়োগ করে উত্তর পাওয়া গেল ২৪৮৪। এই বিয়োগ অঙ্কটি তুমি নিচের বিকল্পগুলোর মধ্য থেকে কোনভাবে লিখবে?

- (ক)
$$\begin{array}{r} ২৬১৩ \\ - ১২৯ \\ \hline ২৪৮৪ \end{array}$$
- (খ)
$$\begin{array}{r} ২৬১৩ \\ - ১২৯ \\ \hline ২৪৮৪ \end{array}$$
- (গ)
$$\begin{array}{r} ২৬১৩ \\ - ১২৯ \\ \hline ২৪৮৪ \end{array}$$

(খ) ২৬১৩

১২৯-

২৪৮৪

(ঙ) উপরের কোনটিই নয়

Item 23 notes:

২৫) খুলনা ঢাকা থেকে পশ্চিম দিকে অবস্থিত, নরসিংদী ঢাকা থেকে পূর্ব দিকে অবস্থিত। সিলেট ঢাকা থেকে আরও পূর্ব দিকে অবস্থিত। এই শহরগুলোর মধ্যে কোনটি সর্বাধিক পশ্চিমে অবস্থিত?

(ক) সিলেট

(খ) ঢাকা

(গ) নরসিংদী

(ঘ) খুলনা

(ঙ) উল্লের দেবার মত যথেষ্ট তথ্য নেই

Item 25 notes:

২৬) রাহাতের কাছে ৪১ টি ডাকটিকিট আছে, রুবিনের কাছে ৩০ টি ডাকটিকিট আছে। গিসার কাছে রাহাত আর রুবিন দু'জনের মিলে যতগুলো ডাকটিকিট তার চেয়ে ২৩ টি ডাকটিকিট বেশী আছে। গিসার কাছে কতটি ডাকটিকিট আছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

(ক) $৪১ + ৩০ - ২৩$

(খ) $৪১ + ৩০ + ২৩$

(গ) $৩০ - ২৩ + ৪১$

(ঘ) $৪১ - ৩০ - ২৩$

(ঙ) উপরের কোনটিই নয়

Item 26 notes:

২৭) নিচের বাণি জায়গায় এই যোগ অঙ্কটি করঃ $80695 + 19512 + 10086$ ।

Item 27 notes:

২৮) নিচের কোন বিকল্পটি $2139 + 251$ এর যোগফলের সবচেয়ে কাছাকাছি?

- (ক) ২৩৫০
- (খ) ২৩৬০
- (গ) ২৪২০
- (ঘ) ২৪০০
- (ঙ) উপরের কোনটিই নয়

Item 28 notes:

২৯) একজন শিক্ষক তার স্কুলের জন্য নতুন আসবাবপত্র কিনেছেন। তিনি দু'টো টেবিল কিনেছেন যেগুলোর প্রতিটির দৈর্ঘ্য ৭ ফুট এবং ওজন ২৫ কিলোগ্রাম। এছাড়াও তিনি ৫ টি বেঞ্চ কিনেছেন যেগুলোর প্রতিটির দৈর্ঘ্য ৫ ফুট এবং ওজন ২০ কিলোগ্রাম। শিক্ষক হিসেব করে দেখলেন আসবাবপত্রগুলোর মোট ওজন ১৫০ কিলোগ্রাম এবং সেগুলোকে স্কুলে নিয়ে যাবার জন্য একটি বাহনের ব্যবস্থা করলেন। নিচের বিকল্পগুলোর মধ্যে কোনটিতে সবচেয়ে সঠিক হিসেব এবং উপযুক্ত পরিবহন মাধ্যম রয়েছে?

- (ক) ওজন $২৫+২৫+২৫+২৫+২০+২০+২০=১৫০$ কিলোগ্রাম; পরিবহনের জন্য রিকশা
 (খ) ওজন $২৫+২৫+২৫+২৫+২০+২০+২০=১৫০$ কিলোগ্রাম; পরিবহনের জন্য ট্রাক
 (গ) ওজন $২৫+২৫+২০+২০+২০+২০+২০=১৫০$ কিলোগ্রাম; পরিবহনের জন্য ট্রাক
 (ঘ) ওজন $২৫+২৫+২০+২০+২০+২০+২০=১৫০$ কিলোগ্রাম; পরিবহনের জন্য বাস
 (ঙ) উপরের কোনটিই নয়

Item 29 notes:

৩১) সুইমিং পুলের পানিতে বানিকটা লবণ থাকে। কলের পানিতে সবচেয়ে কম লবণ থাকে। তবে আটলান্টিক মহাসাগর বা প্রশান্ত মহাসাগরের পানিতে ধূচুর লবণ থাকে। কোন পানিতে সবচেয়ে বেশী লবণ থাকে?

- (ক) কলের পানিতে
 (খ) সুইমিং পুলের পানিতে
 (গ) আটলান্টিক মহাসাগরের পানিতে
 (ঘ) প্রশান্ত মহাসাগরের পানিতে
 (ঙ) উত্তর দেবার মত যথেষ্ট তথ্য নেই

Item 31 notes:

৩২) নতুন ক্লাসের বই-বাতা কেনার জন্য রবি বাসা থেকে ২৫০০ টাকা পেয়েছে। নুসরাত রবির চেয়ে ৭০০ টাকা কম পেয়েছে আর রিমা নুসরাতের চেয়ে ৩০০ টাকা কম পেয়েছে। রিমা কত টাকা পেয়েছে তা হিসেব করার জন্য নিচের কোন গাণিতিক বিবৃতিটি সঠিক হবে?

- (ক) $২৫০০ + ৭০০ + ৩০০$
- (খ) $২৫০০ - ৭০০ + ৩০০$
- (গ) $২৫০০ + ৭০০ - ৩০০$
- (ঘ) $২৫০০ - ৭০০ - ৩০০$
- (ঙ) উপরের কোনটিই নয়

Item 32 notes:

৩৩) নিচের বালি জায়গায় এই বিয়োগ অঙ্কটি করঃ ৪৯১২ - ২৩৮৫ - ১৪৫৬।

Item 33 notes:

৩৪) ৭৩২২ - ৩৯৮০ এর বিয়োগফলের জন্য সঠিক উত্তর নিচের কোনটি?

- (ক) ৩৩৪২
- (খ) ৪৩৪২
- (গ) ৩২৪২
- (ঘ) ৪২৪২
- (ঙ) উপরের কোনটিই নয়

Item 34 notes:

৩৫) একটি স্কুলের প্রধান শিক্ষক স্কুলে বিজয় দিবস উদযাপনের আয়োজন করতে চাইছেন যেখানে সব ছাত্র-ছাত্রীর হাতে বাংলাদেশের একটি করে ছোট পতাকা থাকবে। এই পতাকাগুলো ২০০ সির প্যাকেটে পাওয়া যায়। স্কুলে এমন ৪ টি প্যাকেট আছে। ওই স্কুলে মোট ৬৫৬ জন ছাত্র-ছাত্রী আছে কিন্তু তাদের মধ্যে ১২৬ জন বিজয় দিবসের অনুষ্ঠানে উপস্থিত থাকতে পারবে না। প্রধান শিক্ষক হিসেব করে দেখলেন অনুষ্ঠানে মোট ৫৩০ জন ছাত্র-ছাত্রী থাকবে এবং নতুন করে পতাকাও কিনতে হবে না। নিচের কোন বিকল্পটিতে অনুষ্ঠানে উপস্থিত ছাত্র-ছাত্রী সংখ্যা বের করার জন্য সঠিক গণনা ও উত্তর এক কত প্যাকেট পতাকা ব্যবহার করতে হবে তার সঠিক উত্তর আছে?

- (ক) $৬৫৬ + ১২৬ = ৫৩০$; ৪ প্যাকেট পতাকা
 (খ) $৬৫৬ - ১২৬ = ৫৩০$; ৪ প্যাকেট পতাকা
 (গ) $৬৫৬ - ১২৬ = ৫৩০$; ৩ প্যাকেট পতাকা
 (ঘ) $৬৫৬ - ১২৬ = ৫৩০$; ২ প্যাকেট পতাকা
 (ঙ) উপরের কোনটিই নয়

Item 35 notes:

৩৬) পিয়াকে নিচের অঙ্কটি করতে দেয়া হয়েছিল: 'রাশেল স্কুলে যাবার জন্য বাসা থেকে সকাল ১০.৪০ এ বের হয়ে সকাল ১১.৩০ এ ক্লাসে পৌঁছাল। তার ক্লাসে পৌঁছতে কত সময় লেগেছে?' পিয়া অঙ্কটি করে উত্তর পেল ৫০। উত্তরটা সে কিভাবে লিখবে?

- (ক) সকাল ৫০
 (খ) ৫০ ঘণ্টা
 (গ) ৫০ মিনিট
 (ঘ) ০.৫০ ঘণ্টা
 (ঙ) উপরের কোনটিই নয়

Item 37 notes:

৩৮) লিয়নকে নিচের অঙ্কটি করতে দেয়া হয়েছিল: 'জেনি ১২.৫ টাকা দিয়ে ৫ টি চকলেট কিনছে। সে দোকানীকে ২০ টাকার একটা নোট দিল। দোকানী তাকে কত ফেরত দেবে?' লিয়ন অঙ্কটা করে উত্তর পেল ৭.৫। উত্তরটা সে কিভাবে দিবে?

- (ক) ৭ টাকা
- (খ) ৭ টাকা ৫০ পয়সা
- (গ) ৭.৫ পয়সা
- (ঘ) ৭.৫
- (ঙ) উপরের কোনটিই নয়

Item 38 notes:

Additional Questions:

১. এই পরীক্ষার কোন প্রশ্নের উত্তর দিতে কি তুমি সমস্যায় পড়েছ? কোন প্রশ্ন/প্রশ্নগুলো? কি ধরনের সমস্যায় পড়েছ?

Additional Q1 Notes:

২. যোগ বিয়োগ করতে গিয়ে তুমি কী ধরনের কৌশল ব্যবহার কর? এদের মধ্যে কোন কৌশল কি বেশী ব্যবহার কর?

Additional Q2 Notes:

৩. কিসের উপর ভিত্তি করে তুমি কৌশল নির্ধারণ কর?

Additional Q3 Notes:

৪. তুমি ২৫+৩৭ এই যোগটি কিভাবে করবে? ধাপে ধাপে আমাকে বল।

Additional Q4 Notes:

৫. তুমি ৫৪-২৮ এই বিয়োগটি কিভাবে করবে? ধাপে ধাপে আমাকে বল।

Additional Q5 Notes:

৬. কোন একটি গাণিতিক সমস্যার যে উত্তর তুমি পেয়েছ সেটি সঠিক কী না তা তুমি কিভাবে যাচাই কর?

Additional Q6 Notes:

৭. পাটিগণিতের শব্দ-সম্মিলিত সমস্যা বা বাক্যক্কে সমাধান করার সময় তোমার কোন ভুল সবচেয়ে বেশী হয় বলে তুমি মনে কর?

Additional Q7 Notes:

Appendix 3.25: Plain language statement (PLS) and consent form for experts

[PLS]



Assessment of Mathematical Problem Solving: Addition and Subtraction in Arithmetic Word Problem Solving

You have been invited to participate in research being conducted as a doctoral study under the supervision of Professor Mark Wilson and Professor David Clarke of the Melbourne Graduate School of Education at the University of Melbourne.

What are the aims of the research?

This research aims to develop formative assessment techniques which will help teachers to identify specific difficulties students face while solving arithmetic word problems. As a consequence, it is hoped that students will get better instruction in areas where they are struggling.

What is your involvement in the research?

A set of Mathematics items have been developed to assess students of Class 4 and Class 5 for this research. These items ask students to respond to some questions related to arithmetic and to do some calculations on a paper. Your participation in this research will entail giving your opinion about these test items. In addition, some basic demographic information about yourself will be collected. The whole data collection process should take approximately 60 to 90 minutes of your time and should not have any adverse effect for you.

Information collected from you will be used with the highest level of confidentiality. The data will be de-identified and none of the participants will be named at any stage of disseminating the findings. Utmost care will be taken to ensure findings cannot be linked back to individuals.

How will confidentiality be protected?

Your participation in this study is completely voluntary. Should you wish to withdraw yourself or any of your data that has not been processed, you are completely free to do so. The information gathered during the study will be stored in secured files at the University of Melbourne, under the University's guidelines for the management of research data and records, and destroyed five years after publication of written findings.

How do you agree to participate?

If you agree to take part in the study, please indicate that you have read and understood this information by signing the accompanying consent form and returning it to Farhan Azim in scanned format to the email address (fazim@student.unimelb.edu.au).

HREC: 1648228.1; Date: 19/10/20; Version: PLS for opinion survey (teachers and experts)

Melbourne Graduate School of Education
The University of Melbourne Victoria 3010 Australia
T: +61 3 8344 8285 F: +61 3 8344 8529 W: www.edfac.unimelb.edu.au

How do you obtain further information?

Should you require any further information or have any concerns, please do not hesitate to contact Professor Mark Wilson who is the responsible researcher for this study. His email is wilson.m@unimelb.edu.au. If you are in Bangladesh and would rather talk with someone in the country, you can contact Professor Hosne Ara Begum of the University of Dhaka who is working as a mentor in this research project. Her email is hosne.dhaka@gmail.com and her phone number is +8801791500218.

This research project has been approved by the Human Research Ethics Committee of The University of Melbourne. If you have any concerns or complaints about the conduct of this research project, which you do not wish to discuss with the research team, you should contact the Manager, Human Research Ethics, Office for Research Ethics and Integrity, University of Melbourne, VIC 3010. Tel: +61 3 8344 2073 or Fax: +61 3 9347 6739 or Email: HumanEthics-complaints@unimelb.edu.au. All complaints will be treated confidentially. In any correspondence please provide the name of the research team or the name or ethics ID number of the research project.

Thank you for reading this statement and considering its contents.

Yours sincerely,

Professor Mark Wilson
Melbourne Graduate School of Education
The University of Melbourne
email: wilson.m@unimelb.edu.au

Professor David Clarke
Melbourne Graduate School of Education
The University of Melbourne
email: d.clarke@unimelb.edu.au

Farhan Azim
PhD Student
Melbourne Graduate School of Education
The University of Melbourne
email: fazim@student.unimelb.edu.au

[Consent form]



**Assessment of Mathematical Problem Solving: Addition and Subtraction in
Arithmetic Word Problem Solving**

Name of participant:

Name of institution:

Name of Investigators: Prof. Mark Wilson, Prof. David Clarke, Farhan Azim

I consent to participate in the above research. I understand the nature of the research and what is expected of me and agree my participation on that basis.

I acknowledge that:

- I have received a copy of and read the Plain Language Statement explaining the study;
- I have received an adequate explanation of all likely risks, effects, discomforts or inconvenience arising from participation in the project;
- I understand that participation in the study is voluntary, that I have the right to withdraw at any time, and that I may withdraw any data supplied up to the point of analysis and publication;
- I understand that I will participate in an opinion survey;
- I understand that some demographic information about me will be collected;
- I have been informed that the confidentiality of the information I provide will be safeguarded, subject to any legal requirements (e.g., subpoena, freedom of information, mandated reporting), and in the following ways:
 - no names or personal details of individual participants will be revealed in any report of the study, and any contextual details that might reveal their identity will be removed;
 - only the investigators will have access to the data collected in this study; and
 - data will be stored securely in a locked facility in the Assessment Research Centre, Melbourne Graduate School of Education, and disposed-off securely five years after the completion of the written findings of the research.

Signature of Participant:..... **Date:**

HREC: 1648228.1; Date: 19/10/20; Version: Consent form for teachers and experts

Melbourne Graduate School of Education
The University of Melbourne Victoria 3010 Australia
T: +61 3 8344 8285 F: +61 3 8344 8529 W: www.edfac.unimelb.edu.au

Appendix 3.26: Codebook for the data collection in the final phase

SchRID				
		Value	Count	Percent
Standard Attributes	Label	School		
		Research ID		
	Type	Numeric		
	Measurement	Nominal		
Valid Values	2		62	7.7%
	3		121	15.0%
	4		85	10.5%
	5		64	7.9%
	6		107	13.2%
	7		145	17.9%
	8		89	11.0%
	9		89	11.0%
	10		46	5.7%

StudSID		
		Value
Standard Attributes	Label	Student School
		ID
	Type	Numeric
	Measurement	Nominal

StudRID		
		Value
Standard Attributes	Label	Student
		Research ID
	Type	Numeric
	Measurement	Nominal

StudGrade

		Value	Count	Percent
Standard Attributes	Label	Student's Class		
	Type	Numeric		
	Measurement	Nominal		
Valid Values	4		412	51.0%
	5		396	49.0%

StudClaSec

		Value	Count	Percent
Standard Attributes	Label	Student's Class Section		
	Type	Numeric		
	Measurement	Scale		
N	Valid	808		
	Missing	0		
Labeled Values	0	None	359	44.4%
	1	Section A	191	23.6%
	2	Section B	208	25.7%
	3	Section C	50	6.2%
	4	Section D	0	0.0%

StudGender

		Value	Count	Percent
Standard Attributes	Label	Student's Gender		
	Type	Numeric		
	Measurement	Nominal		
Valid Values	1	Female	474	58.7%
	2	Male	331	41.0%
Missing Values	System		3	0.4%

StudBirthYear

		Value
Standard Attributes	Label	Student's Year of Birth
	Type	Numeric
	Measurement	Scale
	Valid	755
N	Missing	53

StudPrePrim

		Value	Count	Percent
Standard Attributes	Label	If Student Attended Pre-primary		
	Type	Numeric		
	Measurement	Scale		
	Valid	799		
N	Missing	9		
Labeled Values	1	Attended Pre-Primary	519	64.2%
	2	Did not attend Pre-Primary	280	34.7%

StudPrivTutio

		Value	Count	Percent
Standard Attributes	Label	If Student Takes Private Tuition		
	Type	Numeric		
	Measurement	Nominal		
Valid Values	1	Takes private tuition	505	62.5%
	2	Does not take private tuition	293	36.3%
Missing Values	System		10	1.2%

ParentEdMo

		Value	Count	Percent
Standard Attributes	Label	Mother's level of education (in years)		
	Type	Numeric		
	Measurement	Nominal		
Valid Values	0		10	1.2%
	1		3	0.4%
	2		7	0.9%
	3		13	1.6%
	4		21	2.6%
	5		113	14.0%
	6		16	2.0%
	7		17	2.1%
	8		56	6.9%
	9		28	3.5%
	10		92	11.4%
	11		2	0.2%
	12		30	3.7%
	15		23	2.8%
	17		4	0.5%
Missing Values	System		373	46.2%

ParentEdFa

		Value	Count	Percent
Standard Attributes	Label	Father's level of ed (in years)		
	Type	Numeric		
	Measurement	Nominal		
Valid Values	0		12	1.5%
	1		6	0.7%
	2		5	0.6%
	3		10	1.2%
	4		19	2.4%
	5		86	10.6%
	6		22	2.7%
	7		22	2.7%
	8		34	4.2%
	9		27	3.3%
	10		105	13.0%
	11		1	0.1%

	12		22	2.7%
	15		45	5.6%
	17		6	0.7%
Missing Values	System		386	47.8%

ParentProfMo

		Value	Count	Percent
Standard Attributes	Label	Mother's occupation		
	Type	String		
	Measurement	Nominal		
Valid Values			189	23.4%
	ATHLETE		1	0.1%
	BUSINESS		1	0.1%
	GARMENT WORKER		15	1.9%
	GARMENTS WORKER		48	5.9%
	HOSUEWIFE		3	0.4%
	HOUSEMAID		1	0.1%
	Housewife		11	1.4%
	HOUSEWIFE		386	47.8%
	LABOUR		41	5.1%
	SERVICE		99	12.3%
	TAILOR		8	1.0%
	Teacher		1	0.1%
	TEACHER		2	0.2%
	UNEMPLOYED		1	0.1%
	WORKS ABROAD		1	0.1%

ParentProfFa

		Value	Count	Percent
Standard Attributes	Label	Father's occupation		
	Type	String		
	Measurement	Nominal		
Valid Values			188	23.3%
	ARTIST		1	0.1%
	BANKER		1	0.1%
	BARBER		3	0.4%
	BLACKSMITH		1	0.1%

Business	1	0.1%
BUSINESS	136	16.8%
CARPENTER	1	0.1%
CHEF	2	0.2%
COBBLER	1	0.1%
DRIVER	60	7.4%
EXPAT	1	0.1%
Farmer	1	0.1%
FARMER	3	0.4%
FOREIGN WORKER	3	0.4%
GARMENT WORKER	2	0.2%
GARMENTS WORKER	28	3.5%
GOVT. SERVICE	2	0.2%
HAWKER	1	0.1%
LABOUR	59	7.3%
MASON	32	4.0%
MOTOR MECHANIC	1	0.1%
PAINTER	1	0.1%
Physician	1	0.1%
PHYSICIAN	4	0.5%
RICKSHAW PULLER	11	1.4%
Service	8	1.0%
SERVICE	239	29.6%
SERVIICE	1	0.1%
Tailor	1	0.1%
TAILOR	9	1.1%
TEACHER	3	0.4%
UNEMPLOYED	2	0.2%

AvFamExpen

		Value
Standard Attributes	Label	Average family expenditure per month
	Type	Numeric
	Measurement	Scale
	Valid	450
	Missing	358

Codes for items

Item_1

		Value	Count	Percent
Standard Attributes	Label	Item 01		
	Type	String		
	Measurement	Nominal		
Valid Values	99		1	0.1%
	A		228	28.2%
	B		5	0.6%
	C		568	70.3%
	D		4	0.5%
	E		2	0.2%

Item_2

		Value	Count	Percent
Standard Attributes	Label	Item 02		
	Type	String		
	Measurement	Nominal		
Valid Values	99		18	2.2%
	A		171	21.2%
	B		22	2.7%
	C		77	9.5%
	D		490	60.6%
	E		30	3.7%

Item_3

		Value	Count	Percent
Standard Attributes	Label	Item 03		
	Type	String		
	Measurement	Nominal		
Valid Values	99		7	0.9%
	A		722	89.4%
	B		74	9.2%
	C		1	0.1%
	D		4	0.5%

Item_4

		Value	Count	Percent
Standard Attributes	Label	Item 04		
	Type	String		
	Measurement	Nominal		
Valid Values	99		11	1.4%
	A		10	1.2%
	B		707	87.5%
	C		23	2.8%
	D		40	5.0%
	E		17	2.1%

Item_5

		Value	Count	Percent
Standard Attributes	Label	Item 05		
	Type	String		
	Measurement	Nominal		
Valid Values	99		16	2.0%
	A		176	21.8%
	B		66	8.2%
	C		29	3.6%
	D		506	62.6%
	E		15	1.9%

Item_6

		Value	Count	Percent
Standard Attributes	Label	Item 06		
	Type	String		
	Measurement	Nominal		
Valid Values	99		10	1.2%
	A		40	5.0%
	B		692	85.6%
	C		21	2.6%
	D		26	3.2%
	E		19	2.4%

Item_7

		Value	Count	Percent
Standard Attributes	Label	Item 07		
	Type	String		

	Measurement	Nominal		
Valid Values	99		7	0.9%
	A		105	13.0%
	B		29	3.6%
	C		22	2.7%
	D		565	69.9%
	E		80	9.9%

Item_8

		Value	Count	Percent
Standard Attributes	Label	Item 08		
	Type	String		
	Measurement	Nominal		
Valid Values	99		32	4.0%
	A		109	13.5%
	B		333	41.2%
	C		15	1.9%
	D		25	3.1%
	E		294	36.4%

Item_9

		Value	Count	Percent
Standard Attributes	Label	Item 09		
	Type	String		
	Measurement	Nominal		
Valid Values	99		24	3.0%
	A		652	80.7%
	B		111	13.7%
	C		12	1.5%
	D		9	1.1%

Item_10

		Value	Count	Percent
Standard Attributes	Label	Item 10		
	Type	String		
	Measurement	Nominal		
Valid Values	99		12	1.5%
	A		620	76.7%

	B		62	7.7%
	C		41	5.1%
	D		19	2.4%
	E		54	6.7%

Item_11

		Value	Count	Percent
Standard Attributes	Label	Item 11		
	Type	String		
	Measurement	Nominal		
Valid Values	99		11	1.4%
	A		64	7.9%
	B		637	78.8%
	C		48	5.9%
	D		31	3.8%
	E		17	2.1%

Item_12

		Value	Count	Percent
Standard Attributes	Label	Item 12		
	Type	String		
	Measurement	Nominal		
Valid Values	99		10	1.2%
	A		147	18.2%
	B		78	9.7%
	C		25	3.1%
	D		502	62.1%
	E		46	5.7%

Item_13

		Value	Count	Percent
Standard Attributes	Label	Item 13		
	Type	String		
	Measurement	Nominal		
Valid Values	99		17	2.1%
	A		167	20.7%
	B		190	23.5%
	C		116	14.4%

	D		177	21.9%
	E		141	17.5%

Item_14

		Value	Count	Percent
Standard Attributes	Label	Item 14		
	Type	String		
	Measurement	Nominal		
Valid Values	99		28	3.5%
	A		124	15.3%
	B		66	8.2%
	C		219	27.1%
	D		249	30.8%
	E		122	15.1%

Item_15

		Value	Count	Percent
Standard Attributes	Label	Item 15		
	Type	String		
	Measurement	Nominal		
Valid Values	99		11	1.4%
	A		651	80.6%
	B		135	16.7%
	C		3	0.4%
	D		8	1.0%

Item_16

		Value	Count	Percent
Standard Attributes	Label	Item 16		
	Type	String		
	Measurement	Nominal		
Valid Values	99		16	2.0%
	A		44	5.4%
	B		268	33.2%
	C		255	31.6%
	D		24	3.0%
	E		201	24.9%

Item_17

		Value	Count	Percent
Standard Attributes	Label	Item 17		
	Type	String		
	Measurement	Nominal		
Valid Values	99		13	1.6%
	A		71	8.8%
	B		133	16.5%
	C		505	62.5%
	D		37	4.6%
	E		49	6.1%

Item_18

		Value	Count	Percent
Standard Attributes	Label	Item 18		
	Type	String		
	Measurement	Nominal		
Valid Values	99		20	2.5%
	A		106	13.1%
	B		67	8.3%
	C		117	14.5%
	D		51	6.3%
	E		447	55.3%

Item_19

		Value	Count	Percent
Standard Attributes	Label	Item 19		
	Type	String		
	Measurement	Nominal		
Valid Values	99		10	1.2%
	A		41	5.1%
	B		183	22.6%
	C		31	3.8%
	D		506	62.6%
	E		37	4.6%

Item_20

		Value	Count	Percent
Standard Attributes	Label	Item 20		

Valid Values	Type	String		
	Measurement	Nominal		
	99		28	3.5%
	A		352	43.6%
	B		47	5.8%
	C		51	6.3%
	D		305	37.7%
	E		25	3.1%

Item_21

		Value	Count	Percent
Standard Attributes	Label	Item 21		
	Type	String		
	Measurement	Nominal		
Valid Values	99		15	1.9%
	A		491	60.8%
	B		278	34.4%
	C		16	2.0%
	D		8	1.0%

Item_22

		Value	Count	Percent
Standard Attributes	Label	Item 22		
	Type	String		
	Measurement	Nominal		
Valid Values	99		13	1.6%
	A		242	30.0%
	B		99	12.3%
	C		38	4.7%
	D		238	29.5%
	E		178	22.0%

Item_23

		Value	Count	Percent
Standard Attributes	Label	Item 23		
	Type	String		
	Measurement	Nominal		

Valid Values	99		20	2.5%
	A		411	50.9%
	B		38	4.7%
	C		170	21.0%
	D		71	8.8%
	E		98	12.1%

Item_24

		Value	Count	Percent
Standard Attributes	Label	Item 24		
	Type	String		
	Measurement	Nominal		
Valid Values	99		25	3.1%
	A		95	11.8%
	B		357	44.2%
	C		83	10.3%
	D		108	13.4%
	E		140	17.3%

Item_25

		Value	Count	Percent
Standard Attributes	Label	Item 25		
	Type	String		
	Measurement	Nominal		
Valid Values	99		19	2.4%
	A		73	9.0%
	B		125	15.5%
	C		76	9.4%
	D		462	57.2%
	E		53	6.6%

Item_26

		Value	Count	Percent
Standard Attributes	Label	Item 26		
	Type	String		
	Measurement	Nominal		
Valid Values	99		31	3.8%
	A		264	32.7%
	B		342	42.3%

	C		31	3.8%
	D		89	11.0%
	E		51	6.3%

Item_27

		Value	Count	Percent
Standard Attributes	Label	Item 27		
	Type	String		
	Measurement	Nominal		
Valid Values	99		21	2.6%
	A		549	67.9%
	B		220	27.2%
	D		18	2.2%

Item_28

		Value	Count	Percent
Standard Attributes	Label	Item 28		
	Type	String		
	Measurement	Nominal		
Valid Values	99		26	3.2%
	A		62	7.7%
	B		268	33.2%
	C		57	7.1%
	D		130	16.1%
	E		265	32.8%

Item_29

		Value	Count	Percent
Standard Attributes	Label	Item 29		
	Type	String		
	Measurement	Nominal		
Valid Values	99		26	3.2%
	A		72	8.9%
	B		168	20.8%
	C		327	40.5%
	D		82	10.1%

E		133	16.5%

Item_30

		Value	Count	Percent
Standard Attributes	Label	Item 30		
	Type	String		
	Measurement	Nominal		
Valid Values	99		32	4.0%
	A		48	5.9%
	B		415	51.4%
	C		75	9.3%
	D		67	8.3%
	E		171	21.2%

Item_31

		Value	Count	Percent
Standard Attributes	Label	Item 31		
	Type	String		
	Measurement	Nominal		
Valid Values	99		24	3.0%
	A		39	4.8%
	B		37	4.6%
	C		302	37.4%
	D		335	41.5%
	E		71	8.8%

Item_32

		Value	Count	Percent
Standard Attributes	Label	Item 32		
	Type	String		
	Measurement	Nominal		
Valid Values	99		35	4.3%
	A		213	26.4%
	B		91	11.3%
	C		117	14.5%
	D		307	38.0%
	E		45	5.6%

Item_33

		Value	Count	Percent
Standard Attributes	Label	Item 33		
	Type	String		
	Measurement	Nominal		
Valid Values	99		47	5.8%
	A		141	17.5%
	B		547	67.7%
	C		65	8.0%
	D		8	1.0%

Item_34

		Value	Count	Percent
Standard Attributes	Label	Item 34		
	Type	String		
	Measurement	Nominal		
Valid Values	99		33	4.1%
	A		450	55.7%
	B		112	13.9%
	C		54	6.7%
	D		49	6.1%
	E		110	13.6%

Item_35

		Value	Count	Percent
Standard Attributes	Label	Item 35		
	Type	String		
	Measurement	Nominal		
Valid Values	99		46	5.7%
	A		158	19.6%
	B		192	23.8%
	C		166	20.5%
	D		67	8.3%
	E		179	22.2%

Item_36

		Value	Count	Percent
Standard Attributes	Label	Item 36		

Valid Values	Type	String		
	Measurement	Nominal		
	99		43	5.3%
	A		201	24.9%
	B		139	17.2%
	C		165	20.4%
	D		98	12.1%
	E		162	20.0%

Item_37

		Value	Count	Percent
Valid Values	Standard Attributes	Item 37		
	Type	String		
	Measurement	Nominal		
	99		38	4.7%
	A		100	12.4%
	B		70	8.7%
	C		408	50.5%
	D		117	14.5%
	E		75	9.3%

Item_38

		Value	Count	Percent
Valid Values	Standard Attributes	Item 38		
	Type	String		
	Measurement	Nominal		
	99		39	4.8%
	A		47	5.8%
	B		335	41.5%
	C		166	20.5%
	D		143	17.7%
	E		78	9.7%

Appendix 4.1: WLE for the pilot study respondents (sorted by fit)

The following table presents WLE for respondents (Sorted by fit) as calculated in *ConQuest*. The PID here (the first column) is the Person ID assigned by *ConQuest* while the analysis is run in the software.

PID	Raw score	WLE	Standard error	Fit
47	24	1.793206581	0.468699225	0.405250539
56	24	1.793206581	0.468699225	0.405250539
21	10	-0.909405816	0.473877098	0.408237085
37	20	1.002677677	0.435624118	0.479688227
52	24	1.793206581	0.468699225	0.510196229
28	13	-0.289840473	0.44453739	0.516602902
7	20	1.002677677	0.435624118	0.521937224
14	13	-0.289840473	0.44453739	0.541824467
59	18	0.635666949	0.430493945	0.550848991
22	14	-0.098397316	0.438533334	0.552209177
53	9	-1.137991324	0.488009218	0.552407321
3	22	1.384016051	0.447605257	0.562521561
20	17	0.454110783	0.430223061	0.570657424
16	15	0.088428389	0.434200013	0.572598715
41	19	0.818088591	0.432274433	0.58254685
45	23	1.584078353	0.456777479	0.598155996
23	16	0.272135843	0.431449663	0.603239858
43	12	-0.487551869	0.452331619	0.622849416
49	16	0.272135843	0.431449663	0.65910551
11	16	0.272135843	0.431449663	0.663895265
9	11	-0.693379901	0.462054619	0.690960372
13	16	0.272135843	0.431449663	0.72222585
27	19	0.818088591	0.432274433	0.73032846
48	20	1.002677677	0.435624118	0.730555115
50	19	0.818088591	0.432274433	0.743739192
12	16	0.272135843	0.431449663	0.760422296
4	16	0.272135843	0.431449663	0.772245964
51	23	1.584078353	0.456777479	0.784113911
15	15	0.088428389	0.434200013	0.836644755
6	14	-0.098397316	0.438533334	0.848367922
39	16	0.272135843	0.431449663	0.854424053
19	18	0.635666949	0.430493945	0.884936675
54	13	-0.289840473	0.44453739	0.893425496
38	20	1.002677677	0.435624118	0.900087562
62	17	0.454110783	0.430223061	0.900956287
17	12	-0.487551869	0.452331619	0.908980963
8	21	1.190811059	0.440664976	0.913528936
10	19	0.818088591	0.432274433	0.924110819
30	17	0.454110783	0.430223061	0.948901456
42	21	1.190811059	0.440664976	1.003747773
35	23	1.584078353	0.456777479	1.009111935
44	24	1.793206581	0.468699225	1.009805635
32	21	1.190811059	0.440664976	1.010510601
2	26	2.251436194	0.504331022	1.036092409
24	13	-0.289840473	0.44453739	1.076649455

57	22	1.384016051	0.447605257	1.147892424
5	27	2.510553312	0.531487866	1.182543587
1	27	2.510553312	0.531487866	1.220072358
58	14	-0.098397316	0.438533334	1.222751278
61	12	-0.487551869	0.452331619	1.229043308
25	13	-0.289840473	0.44453739	1.232011257
18	12	-0.487551869	0.452331619	1.233072056
40	22	1.384016051	0.447605257	1.265266576
26	9	-1.137991324	0.488009218	1.373725889
46	10	-0.909405816	0.473877098	1.440381134
55	14	-0.098397316	0.438533334	1.447045365
34	20	1.002677677	0.435624118	1.587111055
31	13	-0.289840473	0.44453739	1.653557692
36	20	1.002677677	0.435624118	1.707110815
33	18	0.635666949	0.430493945	1.785126939
60	8	-1.381834443	0.50484546	2.197185068
29	11	-0.693379901	0.462054619	2.493468435

Appendix 6.1: Ability groups of cognitive interview respondents

Students who had a merit positions within 1 to 5 were considered to be high ability, students with merit positions 6 to 15 were considered to be medium ability and students with merit positions 16 and beyond were considered to be low ability students.

No.	Student Code	School RID	Student's Grade	Student's Gender	Perceived Ability Group
01	34101_F	3	4	Girl	High
02	34102_F	3	4	Girl	High
03	34107_M	3	4	Boy	Medium
04	34245_M	3	4	Boy	Low
05	35001_M	3	5	Boy	High
06	35003_M	3	5	Boy	High
07	35006_F	3	5	Girl	Medium
08	35008_F	3	5	Girl	Medium
09	25001_M	2	5	Boy	High
10	25003_M	2	5	Boy	High
11	25011_M	2	5	Boy	Medium
12	25031_M	2	5	Boy	Low
13	94301_M	9	4	Boy	High
14	94306_F	9	4	Girl	Medium
15	94310_F	9	4	Girl	Medium
16	94366_F	9	4	Girl	Low
17	95108_F	9	5	Girl	Medium
18	95203_F	9	5	Girl	High
19	95224_F	9	5	Girl	Low
20	95230_F	9	5	Girl	Low

High ability group = 8 students

Medium ability group = 7 students

Low ability group = 5 students