Author/s:
Kamei, T;Pavlovic, M

Title:
Investigating differential item functioning to validate a thinking skills learning progression for students with intellectual disability and autism spectrum disorder

Date:
2021-01-01

Citation:
Kamei, T. & Pavlovic, M. (2021). Investigating differential item functioning to validate a thinking skills learning progression for students with intellectual disability and autism spectrum disorder. International Journal of Educational Research, 106, https://doi.org/10.1016/j.ijer.2020.101726.

Persistent Link:
https://hdl.handle.net/11343/258516

Investigating differential item functioning to validate a thinking skills learning progression for students with intellectual disability and autism spectrum disorder

Toshiko Kamei[a]*, M. Pavlovic[b]

[a]*Assessment Research Centre, Melbourne Graduate School of Education, The University of Melbourne, Melbourne, Australia;* [b]*Assessment Research Centre, Melbourne Graduate School of Education, The University of Melbourne, Melbourne, Australia*

*email: tkamei@student.unimelb.edu.au

Notes on contributors

Toshiko Kamei is a Research Fellow at the Assessment Research Centre, Melbourne Graduate School of Education, The University of Melbourne. Her research interests focus on thinking skills, assessment, and supporting teachers to increase their capacity to teach all students.
PO Box 93, Warragul, VIC Australia 3820
M: +61420309453 E: kameit@unimelb.edu.au
ORCiD: orcid.org/0000-0001-5149-1506

Masa Pavlovic is an Honorary Research Fellow at the Assessment Research Centre, Melbourne Graduate School of Education, The University of Melbourne. Her research interest focus on validity and use of measurement in classroom assessment.
Level 8, 100 Leicester Street
The University of Melbourne, Victoria 3010 Australia
E: m.pavlovic@unimelb.edu.au

ORCiD: orcid.org/0000-0001-7615-8689

The study was conducted with support from the Australian Research Council (LPI140100163), in collaboration with the Victorian Department of Education and Training.

Investigating differential item functioning to validate a thinking skills learning progression for students with intellectual disability and autism spectrum disorder

Thinking skills are recognised as vital components of improved academic and life outcomes. This study's purpose was to develop an assessment of thinking skills for students with intellectual disabilities (ID) and autism spectrum disorder (ASD). A pool of items was developed based on review of research combined with expert judgement. Items were trialled with teachers of 864 students. Within this sample, 616 students were diagnosed with ID, of whom 301 had ASD. Differential item functioning, as part of a partial credit item response model analysis, was used to validate a skills progression for these groups of students. Implications include arguments for teachers to provide instruction in thinking skills to students with less focus on a label or diagnosis but ability.

Keywords: thinking skills; differential item functioning; students with disability; learning progressions

1. **Introduction**

As general capabilities such as thinking skills increasingly become recognised as vital for success in both school and life, (Organization for Economic Co-operation and Development, 2018; Rychen & Salganik, 2003) implementation in international curricula, assessment, and instructional strategies has been picking up pace (Darling-Hammond, 2012). Systems are working to innovate on ways to effectively teach and recognise such skills and assessment is a core component that can affect this transition.

In the study described in this paper, the concept of thinking was based on a definition put forth by Dewey (1910) to define thinking in terms of skills that are observable, measurable, and targeted to learning. Dewey's definition provided the basis from which to investigate the development of skills such as causal reasoning (Baillargeon, Li, Gertner, & Wu, 2010), metacognition (Annevirta & Vauras, 2006;

Flavell, Green & Flavell, 1995), theory of mind (Wellman, 2010), executive functioning (Rueda, Checa & Cómbita, 2012), and working memory (Zelazo & Müller, 2010). Such skills were deemed thinking skills targeted to learning in an educational context.

The Sustainable Development Goal 4: Quality Education (United Nations [UN], 2018) has stated the importance of inclusive education that provides access to all students, including students with disability, to the general curriculum. Research has shown that students with disability can benefit from explicit instruction in thinking skills (Diamond & Lee, 2011) and that such skills are foundational to accessing learning (Kistner et al., 2010). However, there is a dearth of research to investigate the assessment and teaching of thinking skills to students with disability (Spencer, Evmenova, Boon, & Hayes-Harris, 2014). One reason has been the lack of prioritisation of such skills in favour of targeting areas of challenge (Gersten, Harris, & Mason, 2017; Hock, Bulgren, & Brasseur-Hock, 2017). For instance, research in learning and instruction for students with autism spectrum disorder (ASD) has tended to focus on communication and social skills, identified challenges for students with ASD (Begeer et al., 2010; Chen et al., 2016). There is a marked lack of research and evidence targeting thinking skills such as problem solving in mathematics instruction (Bae, Chiang, & Hickson, 2015) or scientific thinking skills (Taylor, Rizzo, Hwang, & Hill, 2019).

In addition, historically, assessments measuring cognitive skills of students with disability have generally been used as norm-referenced diagnostic tools for funding or resource purposes (Garcia, 2015). They are targeted to psychologists or clinicians to use for diagnostic purposes and not for teacher use to inform instruction. Thus, the aim of the study described in this paper was to fill this gap in knowledge by developing an assessment for thinking skills based on a learning progression. The purpose would be for teachers to use to inform instruction and to provide information of the emergence of thinking skills in students with disability.

The assessment developed in this study aimed to describe all students, including students with complex needs in terms of what they can do, not what they cannot. In Australia, the setting of the study, the curriculum begins at foundation level that targets students approximately five years of age (Australian Curriculum, Assessment and Reporting Authority, 2020). It assumes that students have developed the skills necessary to carry out the defined curriculum outcomes at foundation level. However, this may not be the case for all students, including those with a disability. Thus, the thinking skills assessment aimed to extend the range of skills targeted in typical curricula. The developed assessment targeted learning from a Piagetian (1954) pre-intentional stage where students are reacting or responding to stimuli, to a stage where students can hypothesise and plan. Thus, the thinking skills targeted by the assessment were positioned as foundational to and developing towards thinking skills required to access the general curriculum. Thus, it aimed not to replace the curriculum, but to extend the description of skills to target and describe the learning of students whose current level of skills or knowledge are not represented by typical curricula.

1.1 Thinking skills and students with intellectual disability (ID) and ASD

This paper will describe a phase in the development of the thinking skills assessment to examine the item invariance between two groups: students with ID only and students with ID and ASD. The purpose of this phase was to examine whether an assessment of thinking skills can be used for all students with disability, including students with ID and ASD.

Efforts to understand ID has been coloured by political and legal agendas, social and ethical attitudes, psychological and educational practice, theory and research

(Colmar, Maxwell, & Miller, 2006). Thus, the accuracy of government research and statistics have suffered (Colmar et al., 2006). For instance, varying IQ cut-off points based on prevailing public policy have historically been used to determine a diagnosis of ID resulting in inconsistency in the size of the population through time (Parmenter, 2011). Moreover, as scientific knowledge increases in complexity, it has become increasingly clear that people with ID are a heterogeneous group (Burack, Russo, Flores, Iarocci, & Zigler, 2012). While acknowledging the heterogeneous nature of this group, the DSM-5 defines ID as manifesting as:

> deficits in general mental abilities such as reasoning, problem solving, planning, abstract thinking, judgement, academic learning, and learning from experience. The deficits result in impairments of adaptive functioning such that the individual fails to meet standards of personal independence and social responsibility in one or more aspects of daily life, including communication, social participation, academic or occupational functioning, and personal independence at home or in community settings." (APA, 2013, p. 31).

A diagnosis of ASD is based on developmental history and presenting behaviour (Bent, Barbaro & Dissanayake, 2017). The most recent version of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (American Psychiatric Association [APA], 2013) provided two diagnostic criteria for ASD. They are difficulties in social-communication and restricted and repetitive behaviours and interests (APA, 2013). However, there is significant heterogeneity within people with ASD from individuals with above-average communication skills to those who are non-verbal (Baron-Cohen, 2017).

Both ID and ASD are classified as neurodevelopmental disorders in the DSM-5 (APA, 2013). Although estimates vary, co-occurrence of ID with an ASD diagnosis is common (APA, 2013; Clark, Barbaro, & Dissanayake, 2017). To account for the heterogeneity of ASD and to control for intellectual ability, the study described in this paper investigated the differential item functioning (DIF) of students divided into two groups: students with a diagnosis of ID only and students with a diagnosis of both ID and ASD. These classifications were based on the diagnostic criteria put forth in the state of Victoria, Australia where the study took place. Despite an abilities-based approach endorsed by the study with less focus on a label or diagnosis, to constrain the sample population and define groups for DIF analysis, the diagnostic categories as set out in Victoria, Australia were used.

Research has been carried out to investigate reasons why children with ASD may develop atypically to children without a diagnosis of ASD. Theories have sought to describe the possibility that students with ASD progress atypically in thinking skills such as theory of mind or executive functioning. For example, the extreme male brain theory (Baron-Cohen, 2010) and theories of deficits in theory of mind and central coherence (Baron-Cohen, 2017; Frith, 2008) seek to explain challenges faced by individuals with ASD. Another theory has been the notion that students with ASD may have pockets of deficits in executive functioning skills such as planning, working memory, or flexibility (Happé, Booth, Charlton, & Hughes, 2006; O'Hearn, Asato, Ordaz, & Luna, 2008; Ziermans, Swaab, Stockmann, de Bruin, & van Rijn, 2017).

Research has tended to focus on apparent differences observed in individuals with an ASD diagnosis. Such a scientific approach maintains a focus on the disorder aspect of ASD (Baron-Cohen, 2017). This approach has been contested through the social model of disability with a view of disability as socially constructed (Oliver, 2013) and fluid based on teacher perceptions (Milton, 2012).

Thus, through an abilities-based approach to describe all students in terms of what they can do, this study strove to empirically explore the question of whether thinking skills manifest differently for students with an ASD diagnosis. This phase of the study aimed to answer the following research questions: 1) Is there a typical progression for thinking skills to support the learning of students with ID? 2) Does the learning progression apply to students with ID who also have ASD?

2. Method

2.1 Approach to assessment design and validation

The assessment was developed as part of a larger program called the Students with Additional Needs (SWANs)/Abilities Based Learning and Education Support (ABLES) program that was carried out in partnership with the Victorian Department of Education and Training (Victorian Department of Education and Training, 2018). The program aimed to develop assessments based on learning progressions in foundational skills targeting the learning of students with disability.

The method used to develop the assessment in this study was based on a process for defining criterion-referenced assessment frameworks (Griffin, 1993, 2007). This method has been applied in previous work to develop criterion referenced frameworks in the form learning progressions to support teacher assessment and instruction of students with disability (Coles-Janess & Griffin, 2009; Roberts & Griffin, 2009; Woods & Griffin, 2013). Criterion-referencing provides a framework to interpret students' performance in terms of what they can do as determined by pre-set quality criteria (Glaser, 1997) which can be organised into levels of increasing skill along a progression (Griffin, 2007). This method combined the work of Glaser (1997) on criterion-referenced interpretation, Vygotsky's (1980) theory of the Zone of Proximal Development and Rasch (1960) logic to develop a scale comparing item difficulty to student ability.

Questions of validity drove the evidence gathering process. The approach to arguments for the validity of the assessment was based on the work of Messick (1989) and Kane (2016) and followed procedures put forth by Wolfe and Smith (2007a, 2007b) and Wilson (2005). The study took a unified view of validity with construct validity being the unifying element (Messick, 1989). The purpose was to develop an assessment based on a learning progression to provide teachers with information about the level within which a student was working and give information about what a student should learn next to progress in thinking skills.

2.2 Assessment Development

2.2.1 Review of research of thinking skills

A key assumption underlying instrument design was that there was one, unidimensional construct to be measured (Wilson, 2005). The defining of the construct was guided by the aim to describe thinking in terms of something that could be measured and was applicable in an educational context. An investigation of international curricula indicated the enduring influence of Dewey in the way thinking skills are represented in education. Thus, Dewey (1910,1916) provided the basis from which to investigate research to develop a definition for the construct of thinking skills.

Dewey (1910) described three strands of purposeful thinking: beliefs, imagination and reflection. These strands provided the basis from which to conduct a research review. The review resulted in the defining of five capabilities that fit within the three strands. These were *representing*, the ability to observe and organise information (Siegler, DeLoache, Eisenberg, Saffran, & Leaper, 2014), *creating*, the ability to think beyond immediate knowledge (Bijvoet-van & Hoicka, 2014) and *planning*, the ability to think ahead and keep a strategy in mind (Berger, Chin, Basra, &

Kim, 2015). The other two capabilities were *reasoning*, the ability to recognise causal relations (Koslowski & Masnick, 2010) and to understand thinking and its role in other people's actions (Wellman, 2010), and *evaluating*, the ability to judge strategies and outcomes (Annevirta & Vauras, 2006). Based on the review of research, these five capabilities were determined as key skills that make up each of the strands and vital for learning in an educational context.

The research was then consulted to break the five capabilities down into a pool of items and criteria. 18 items were written to describe behaviours that indicate a student's ability to demonstrate a capability. The 18 items were then described in terms of quality criteria that describe how well a student could demonstrate each item. Two to four quality criteria were written to describe increasing proficiency within each item. The method described above to break down the construct into increasingly detailed levels of manifestation was applied as it helps to ensure that the evidence gathered by the items are related as closely as possible to the definition of the construct being measured (Wilson, 2005).

2.2.2 Review by Subject Matter Experts

The thinking skills construct, strands, capabilities, items and quality criteria were presented to 12 subject matter experts (SMEs) to review. SMEs were leading teachers of students with disability with interest in the teaching and learning of thinking skills. The review resulted in a pool of 18 items and 59 quality criteria.

The SMEs were then tasked with the development of 18 rubrics – one for each item. A rubric is a set of quality criteria combined with an item placed on a rating scale (Griffin, Gillis & Calvitto, 2004). Figure 1 shows a set of rubrics for three items that describe behaviours indicative of the capability of planning.

	5.3 Explains or describes a plan or sequence of steps to complete a new and challenging task (e.g. describes the likely steps to complete a practical task)	
4.3 Explains reasons for choosing a preferred or relevant outcome in the context of a familiar task or activity (e.g., choose an indoor activity because it is raining outside)		6.4 Explains the order of multiple steps of a completed familiar task (e.g., describes the sequence of daily tasks in the classroom or in practical activities)
	5.2 Follows multiple step instructions to complete a new or unfamiliar task	6.3 Performs multiple steps of a familiar task in sequence to accomplish a goal (e.g., checks the daily timetable, locates appropriate materials and starts work without supervision or reminders)
		6.2 Attempts different ways to achieve the first step in a familiar task (e.g., attempts different strategies to organise materials for the day such as looking at what others are doing, checking the daily timetable, preparing materials that were used in previous days)
	5.1 Follows a single step instruction provided by a familiar adult to start a new or unfamiliar task or activity (e.g., with support, finds way to a new classroom)	
4.2 Makes a choice between two or more options (e.g. show a preference for one activity or object over another)		6.1 Relies on support (e.g., modelling, verbal or physical prompting) to initiate the first step in a familiar, multi-step task (e.g., unpacking a bag, morning greeting, organising material for an activity)
4.1 Responds to an object, noise or person (e.g., by looking or turning towards them)		
4. Choosing a preferred or relevant outcome	**5. Dividing a new or unfamiliar task into parts or steps**	**6. Sequencing steps to reach a preferred or relevant outcome in a familiar task**

Figure 1. Items within a set of rubrics describing the capability of planning.

Items 4, 5 and 6 are written along the bottom in bold and the quality criteria are ordered above them according to relative difficulty from the bottom to the top. The criteria were numbered according to order of difficulty. For instance, Criterion 4.1 is the least difficult, 4.2 is more difficult and 4.3 is the most difficult criterion within Item 4.

All 18 rubrics were combined side by side, to show progression in relative difficulty from the bottom to the top. The SMEs applied a process of pairwise comparison (Guilford, 1954) to determine the relative difficulty of the criteria. They used their own judgement and referred to stages of development defined by Inhelder and Piaget (1958) and the cognitive dimension of Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001) to judge the relative difficulty of the quality criteria for each item. These combined rubrics provided the foundation for a draft hypothetical learning progression describing learning and development in thinking skills.

2.2.3 Implementation of the large-scale field trial

The rubrics were prepared for a large-scale field trial by writing the items as stems and quality criteria as a set of ordered response multiple choice items, designed to measure students' thinking skills. An example of Item 4 as it appeared online is shown in Figure 2.

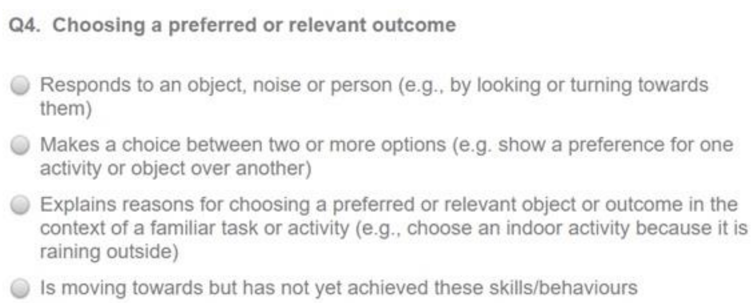

Figure 2. Online format of items.

The 18 items were presented in the form of an online observation assessment that teachers were asked to respond to based on their existing knowledge of their students. Thus, students were not required to respond directly to the items and were indirect participants of the research. One reason for this was that many students with disability may not have the capacity to sit an assessment. In addition, research has shown the value of teacher observational data and the accuracy of their stored knowledge of their students (Hill et al., 2015). As a step to ensure teachers could understand the items and assess students with them, the items were piloted with a group of five SMEs and panelled with four academic researchers with expertise in assessment. This was a final step in preparation for a large-scale field trial of the items.

Participant recruitment

Teachers of students with disability were the direct participants of the study. The definition of disability as set out in the state of Victoria, Australia, the site of the study, was used for the trial. In this study, the term students with disability specifically referred to students who required some level of “reasonable adjustment” (p. 53) to facilitate participation in schools on the same basis as all students (Victoria State Government, 2016). In Victoria, students who receive additional education funding on the grounds of ID must score two standard deviations below a mean score on a standardised test of intelligence (IQ of under 75) (Victoria State Government, 2016). Students with an ASD diagnosis must be recognised to have difficulties in social communication and restricted and repetitive behaviours and interests (APA, 2013; Victoria State Government, 2016).

Teachers were recruited as per the ethics requirements of the University of Melbourne and the Victorian Department of Education and Training. School principals were approached and after obtaining principal consent, teachers were invited to take part in the trial. Teachers were given access to the online observation assessment instrument and asked to assess school age students with disability with the thinking skills observation assessment.

2.3 Differential item functioning

DIF is used to check the reliability of items and investigates the extent that calibrated items are interpreted consistently across different groups (Wolfe & Smith, 2007b). When comparing two groups, one group is referred to as a reference group upon which standard performance is based and the other a focal group. In this study, the reference group was students with ID only and the focal group students with a co-occurrence of ID and ASD. The study followed a DIF analysis process described in Woods and Pavlovic (2016). This process involved investigation of item and item step invariance using a Rasch modelling approach followed by a Mantel Hanszel analysis to interpret the statistical significance of any differences detected.

2.3.1 Method 1: Rasch modelling

The use of Rasch modelling is well suited to investigate for DIF in both dichotomous and polytomous items as item parameters are estimated independently of respondent ability (Osterlind & Everson, 2011). Item Response Theory (IRT) analysis was used to investigate relationships between item difficulty and student ability to achieve the behaviour described in each item criterion. Estimates of item difficulty and student ability were expressed in terms of logits that are calculated according to the logistic response model with the natural log odds for a correct response being:

$$\ln\left[\frac{pv_i}{1-pv_i}\right] = (\theta_v - \delta_i) \tag{1}$$

With θ_v as the ability estimate of person *v* and δ_i is the difficulty estimate of item *i*.

The partial credit approach to IRT modelling (Masters, 1982) was applied as it is suited to polytomous items. A partial score or credit is awarded to each level of performance a respondent obtains. The partial credit model is expressed as:

$$\Pr\,(x_{ni}=\mathrm{x}) = \frac{\exp\sum_{j=0}^{x}(\theta_n-\delta_{ij})}{\sum_{k=0}^{m_i}\exp\,\sum_{j=0}^{k}(\theta_n-\delta_{ij})} \qquad \mathrm{x}=0,\,1,\,\dots\,m_i \tag{2}$$

where the probability of a person *n* scoring *x* on the m_i-step item *i* as a function of the person's position θ_n on the variable and the difficulties of the m_i "steps" in item *i.* The observation *x* is a *count* of the successfully completed item steps, and only the difficulties of these *x completed* steps appear in the numerator of the model.

(Masters, 1982, p. 10, emphasis in original)

Thus, in the Rasch partial credit model, the total score is sufficient to provide both estimates of student ability and item difficulty. An assumption in the Rasch family of models is that item difficulty and person ability parameter estimates are separable and can be estimated independently (Wright & Stone, 1979). Thus, item parameters will remain constant regardless of the group of respondents who are assessed by the items (Masters, 1982). DIF is calculated by comparing items to each other while controlling for overall abilities with the assumption that the assessment is not biased against any particular group (Wu, Tam, & Jen, 2016).

In the partial credit model, estimates of item criteria difficulty are expressed as delta parameters. Deltas are points along the ability scale at which a student has a greater probability of being scored at *k* than *k*-1 (Masters, 1982). It should be noted that the delta parameters depend on the number of students within each criterion and are not independent difficulty parameters, thus it is possible for them to be disordered (Wu et al., 2016). For instance, this can be the case where there are very few students represented by a quality criterion.

In such cases, Thurstonian thresholds are useful. Thresholds represent item criterion difficulty in terms of cumulative probability of achieving a criterion and thus are never disordered. They represent the point at which the cumulative probability of achieving a criterion within an item is 0.5. In other words, they are the point at which a student has at least a 50% chance of achieving at least the indicated level of performance on an item. In this way, they provide cut points to divide the ability continuum into scoring regions within which students have the probability of achieving that criterion or higher.

Item step invariance examines the between-group difference in the difficulty of progressing from category *k*-1 to *k.* This has been termed differential step functioning (Penfield, Myers, & Wolfe, 2008). Step effects are the sum of item and threshold effects thus may not be attributable to item-level properties (Penfield et al., 2008). This means that investigation of item-level effects may not reveal invariance in step-level effects. Thus, in the partial credit model, in consideration of polytomous items, investigation of step-level effects can lead to different conclusions and provide a higher level of sensitivity to between-group variance within items (Penfield et al., 2008).

2.3.2 Method 2: Mantel Haenszel

An assumption in MH analysis is that it is possible to assume that there is a common odds ratio between the reference and focal groups across all the different levels (Wu et al., 2016). The reference group is the source of standard performance on an item in Conquest (Adams, Wu, & Wilson, 2015). The observations are organised according to overall score and performance on an item is compared between the two groups, conditional on score (Zwick, 1990).

The resulting statistic is expressed as differences on the delta scale called the Mantel Haenszel delta difference (MH D-DIF) (Zieky, 1993). In accordance with the partial credit model, the statistic is reported between adjacent score categories. A value of 1.00 indicates that for one group, a category within an item was one delta logit more difficult than for the other group with similar ability (Zieky, 1993). The resulting values have three categories. The first is Category A, indicating negligible or nonsignificant

DIF, which is the case if it is not significant at the 5% level and with an absolute value of less than one. Category B indicates slight to moderate DIF. Category C indicates moderate to large values of DIF (Zieky, 1993), requiring an absolute value significantly greater than one.

2.4 Interpretation of DIF

DIF refers to measurement bias or the fairness of a test and it is investigated to ensure that an assessment instrument is valid for all subgroups of respondents (Wu & Adams, 2007). Its function is to examine whether items function in a similar way for all subgroups while controlling for their abilities. The Standards for Educational and Psychological Testing (American Educational Research Association [AERA], 2014) listed four indications of fairness: 1) equitable treatment of respondents; 2) lack of bias; 3) accessibility; and 4) validity of score interpretations. “A prime threat to fair and valid interpretation of test scores comes from aspects of the test or testing process that may produce construct-irrelevant variance in scores that systematically lowers or raises scores for identifiable groups of test takers” (AERA, 2014, p. 54). DIF is a step to investigate whether there are irrelevant sources of score variation that may affect individuals or groups in a different way (Wise & Plake, 2016). In this way, DIF analysis supports arguments for the generalisability aspect of validity (Wolfe & Smith, 2007b).

DIF is particularly important to investigate in circumstances where theory and prior evidence indicate a possibility of differential prediction based on subgroup (Wise & Plake, 2016). In this case, due to different theories describing the cognitive development of students with ASD (e.g., Baron-Cohen, 2010; Frith, 2008; Happé et al., 2006; O'Hearn et al., 2008; Ziermans et al., 2017), an investigation of students with ASD as a focal group was deemed important.

3. Results

3.1 Results of the large-scale field trial

The thinking skills observation assessment instrument was trialled online by teachers of 864 students with diverse disabilities, including ID and ASD. Of the 864 students, 791 attended specialist settings that specialised in the education of students with disability and 73 attended mainstream settings. The proportion of students in specialist and mainstream settings divided by disability diagnosis is shown in the table below. It should be noted that 82% of the students had more than one disability diagnosis.

Table 1

Proportion of Type of Disability Diagnosis According to Setting

Diagnosis	Students in Mainstream Schools (% of 73 students)	Students in Specialist Schools % of 791 students)	All Students (% of 864 students)
Intellectual disability	28	75	71
Autism Spectrum Disorder	25	62	59
Severe language disorder	20	32	31
Severe behaviour disorder	5	17	16
Physical disability	5	13	13
Visual impairment	5	8	8
Deaf/Hard of hearing	2	4	4

The categories were based on those set out by the Victorian Department of Education and Training (Victoria State Government, 2016). They provided the basis for a set of demographic questions regarding students' disability diagnosis that teachers responded to prior to carrying out the thinking skills assessment. The teachers responded to the questions by choosing "yes", "no", or "I don't know".

The assessment data were analysed using *ConQuest* item response modelling software (Adams et al., 2015). Psychometric analysis of the resulting scale showed good fit to the model and evidence of a unidimensional, underlying construct. Item fit was examined by the weighted mean square fit statistics in terms of the convention set forth by Adams and Khoo (1996) of a lower and upper bound of 0.75 and 1.33. Investigations of item difficulty and person ability showed that the items targeted the ability range of the target population well and there was a good representative spread of students of different ability levels. The reliability indices were also acceptable, with a person separation reliability of 0.97, an item separation reliability index of 0.99 and an alpha coefficient of 0.97. Such results of the calibration of item and person parameters provided evidence that the thinking skills assessment was targeting students with disability as intended with consistency.

Following a review of statistical output with SMEs, a final derived learning progression for thinking skills was established (Figure 3).

7. The student is starting to experiment, adapt, modify, extend, and transfer strategies to new situations and explain the reasoning behind the strategy choice

Increasing Proficiency

6. The student is beginning to explain choices and generalise rules by systematically working through ways to successfully navigate unfamiliar or challenging activities, tasks, or objects

5. The student is beginning to independently use knowledge of successful strategies and outcomes to attempt unfamiliar tasks or activities and engage with new objects.

4. The student is starting to transfer basic familiar strategies to attempt unfamiliar activities or tasks and understand another's knowledge or intention.

3. The student is learning to follow instructions with support and discover causal explanations by recognising whether a strategy or outcome was successful or unsuccessful.

2. The student is starting to carry out trial and error by engaging with obstacles and attempting basic, one-step (even if incorrect) strategies.

1. The student is beginning to respond and react to familiar activities or people.

Figure 3. Empirically derived progression for thinking skills.

The learning progression was derived through ordering item criteria by estimated difficulty and setting cut scores (Wolfe & Smith, 2007a). Through a process of review with SMEs, qualitative transitions were defined by describing groups of criteria clustered at similar levels of difficulty. This resulted in eight levels as shown in Figure 3. The cut scores were expressed as test scores to report on teacher responses to the online observation assessment. Results were presented by reporting where the student was working and where they needed to go next to progress on the thinking skills learning progression.

3.2 DIF analysis

In the response data of 864 students, there were 616 students with ID and within this number, 301 students with an additional diagnosis of ASD. In an IRT context, DIF can be carried out for large-scale assessments. Thus, the sample size should be large, preferably >250 (Embretson & Reise, 2000). The Educational Testing Service recommends a total of 500 and at least 200 for each subgroup (Zwick, 2012). The sample size was deemed to be sufficiently large overall and in each category to carry out a DIF analysis. The data from the two groups: students with ID only (n=315) and students with ASD and ID (n=301) were analysed using Conquest computer software (Adams et al., 2015).

3.2.1 Investigation of bi-variate plots

To check the stability of the items across groups, estimates of difficulty were separately calculated for each group applying the Rasch partial credit model (Masters, 1982). The estimates were compared in scatterplots to visually compare whether there was variation beyond what can be explained by standard error.

A Rasch partial credit model analysis was carried out on 315 students identified as having ID only and for 301 students with ASD and ID. The item parameters of these groups were compared with the item parameters from the original analysis of 864 students to check that the progression derived could be reproduced for these subgroups. In other words, that the items and item criteria maintained their rank ordering and estimates of difficulty were stable. Comparisons of item parameters are shown in Figures 4 and 5.

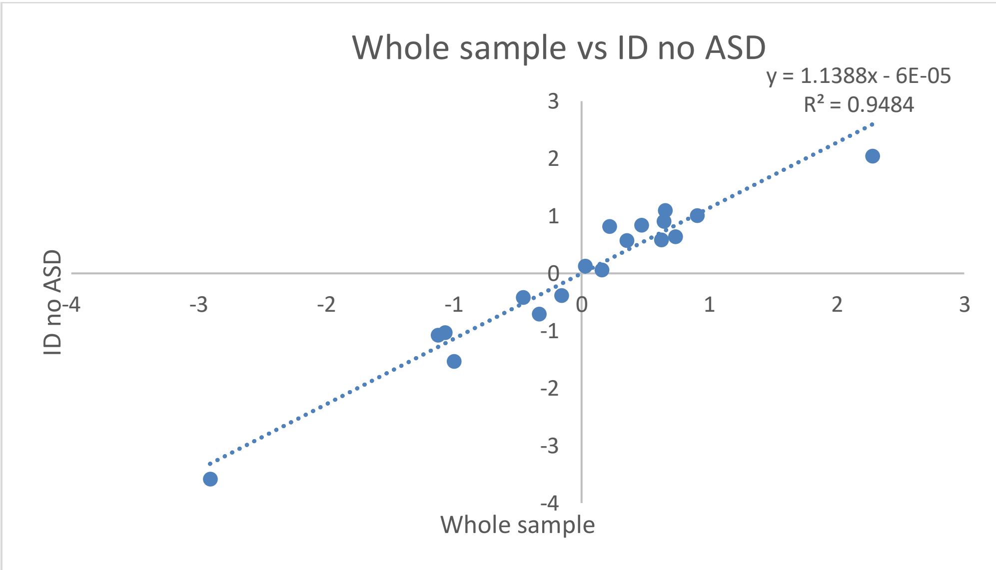

Figure 4. Scatterplot of item parameter estimates of students with ID only compared to the whole sample.

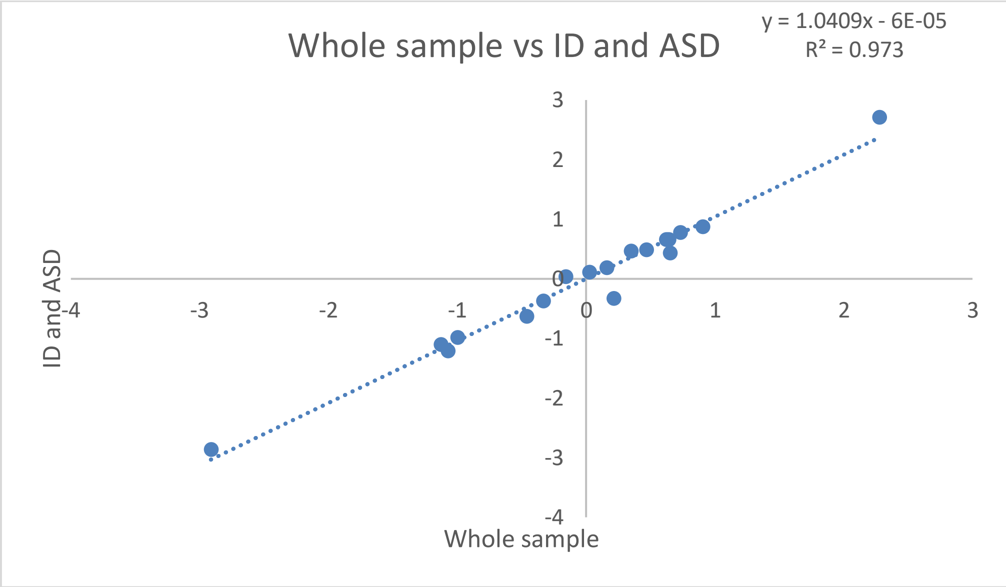

Figure 5. Scatterplot of item parameter estimates of students with ASD and ID compared to the whole sample.

Figure 4 shows item parameter estimates for all students in the calibration sample plotted along the x-axis and for students with ID only on the y-axis. The slope of the line of best fit is very close to one and the correlation between the two sets of parameters is high at 0.95. Figure 5 shows the comparison of all students plotted along the x-axis and students with ID and ASD along the y-axis. This shows a similar pattern with the slope of the line of best fit very close to one and a correlation of 0.97.

Next, the delta difficulty estimates of item criteria were compared. These comparisons are shown in Figures 6 and 7.

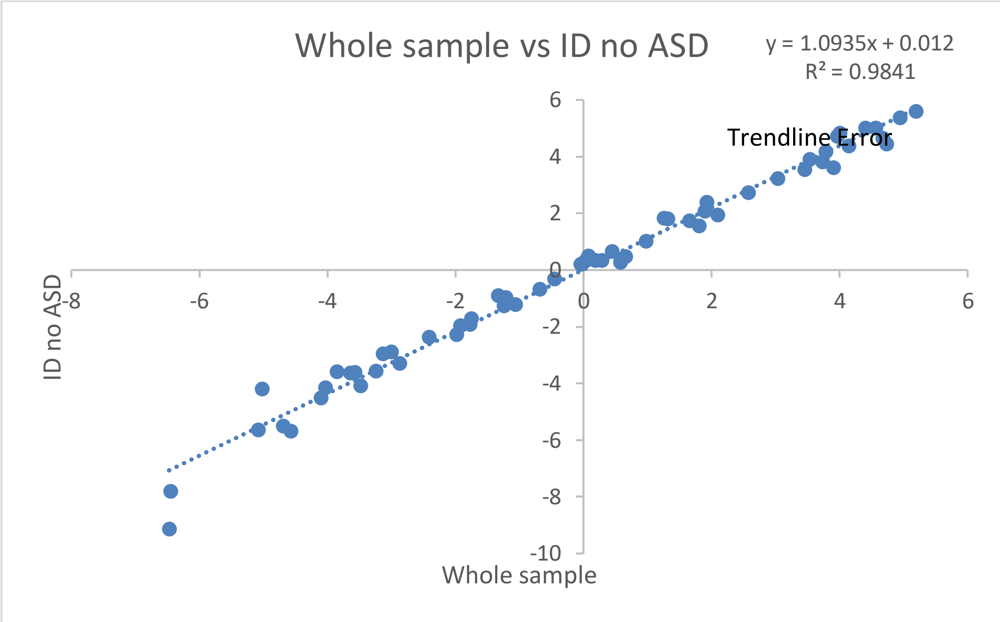

Figure 6. Scatterplot of delta difficulty estimates of students with ID only compared to the whole sample.

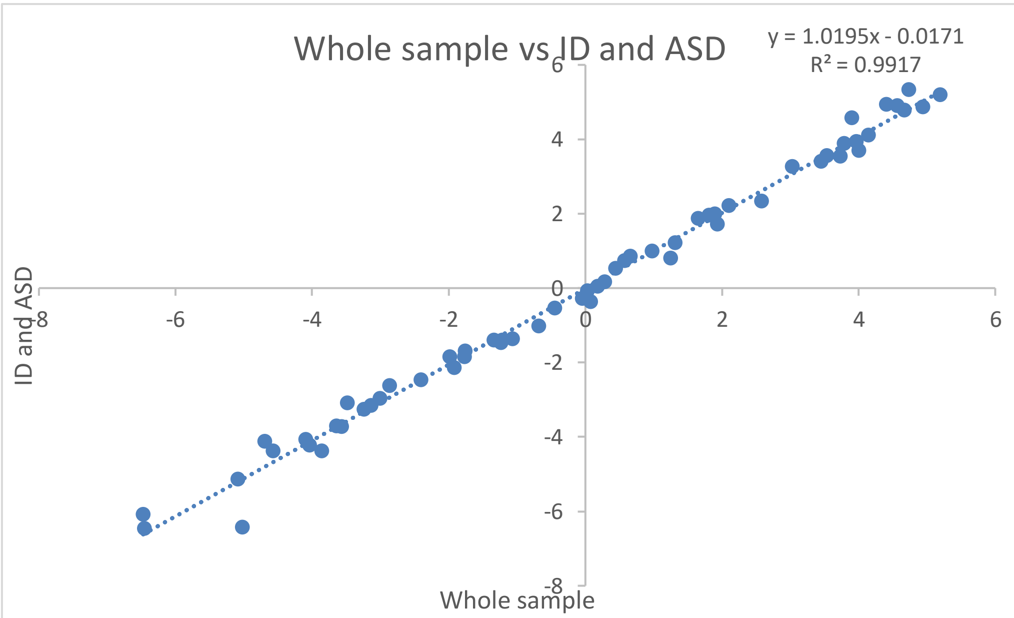

Figure 7. Scatterplot of delta difficulty estimates of students with ASD and ID compared to the whole sample.

Figure 6 shows delta estimates of item criteria for all students plotted in the calibration sample along the x-axis and for students with ID only on the y-axis. The slope of the line of best fit is very close to one and the correlation between the two sets of delta estimates is high at 0.98. Figure 7 also demonstrates a high correlation of 0.99

of the comparison of difficulty thresholds of the whole sample on the x-axis and students with ASD and ID on the y-axis.

The comparisons illustrated in Figures 4 through 7 demonstrated a very high level of similarity, rather than difference, in the hierarchy of thinking skills for students with ID only, ASD and ID and the broader group of students. The comparisons of the difficulty parameters and order of emergence of criteria between the whole sample, students with ID and students with ID and ASD addressed research question number 1. Research question 1 was posed to investigate whether a typical progression for thinking skills exists for students with ID. The results of the comparisons shown in the bi-variate plots provided evidence for this as the criteria maintained their rank ordering and estimates of difficulty were stable in the comparisons. In the next step, the item parameters of students with ID only and students with ID and ASD were compared to investigate research question 2: the extent to which the learning progression was applicable for students with ID and ASD.

Figure 8 shows a comparison of the item parameters for students with ID only (n=315) and students with ID and ASD (n=301).

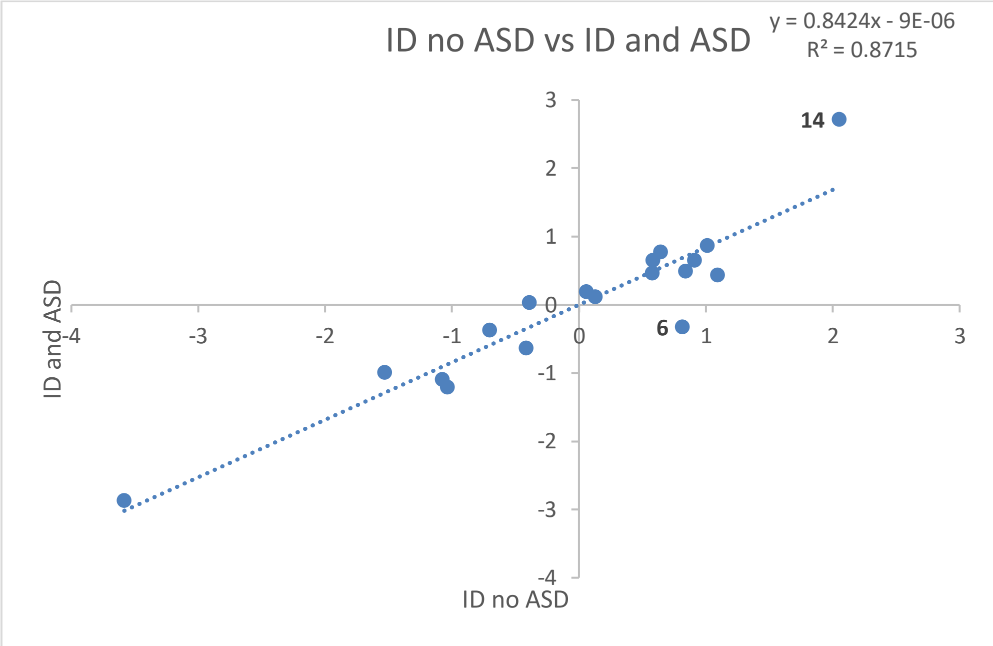

Figure 8. Scatterplot of item parameter estimates of students with ID only compared to students with ID and ASD.

The comparison of item parameters of students with ID only compared to students with ID and ASD demonstrated a high level of similarity. The line of best fit was close to one and the correlation was high at 0.87. Items 6 and 14 were flagged as possible outliers. The standardised difference of Items 6 and 14 were computed by dividing the difference in item parameter estimates by the square root of the sum of squares of the standard errors. This standardised difference can be regarded as a z-statistic, so a value outside of -2 to +2 could be considered statistically significant (Wu et al., 2016). Item 6 had a standardised difference of 6.07 with an absolute difference in logits of 1.14. This meant that Item 6 (Sequencing steps to reach a preferred or relevant outcome in a familiar task) was slightly more difficult for students with ID only than for students with ID and ASD. Item 14 had a standardised difference of -3.02 with an absolute difference in logits of -0.67 This meant Item 14 (Understanding another's viewpoint) was slightly less difficult for students with ID only than for students with ID and ASD. These items were flagged for further investigation using the Mantel-Haenszel method.

Next, the delta difficulty estimates of the item criteria were compared between students with ID only and students with ID and ASD. This is shown in Figure 9.

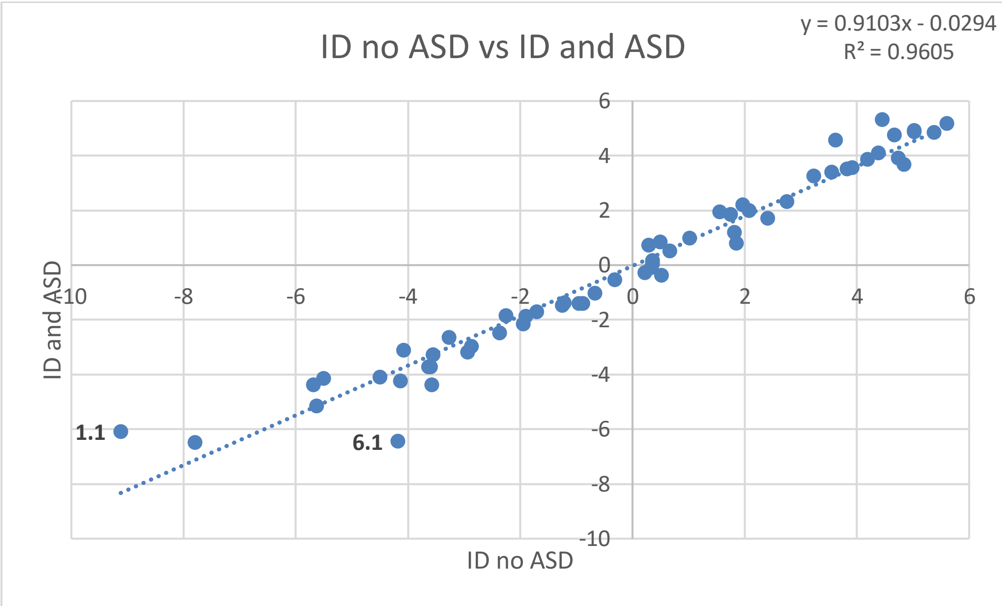

Figure 9. Scatterplot of delta difficulty estimates of students with ID only compared to students with ID and ASD.

The line of best fit was close to one and the correlation was high at 0.96. Criteria 1.1 and 6.1 were flagged as possible outliers. A further investigation of Criteria 1.1 and 6.1 revealed that there were very few students assigned to these criteria. For instance, only 2% of students with ASD were assigned Criterion 1.1 and 1.67% were assigned Criterion 6.1. Thus, these criteria were of very low difficulty and chosen for very few students meaning even small changes in proportions of students in the lowest category groups would inflate the size of detected differences.

The comparisons of item and item criteria parameters showed that the estimates of difficulty maintained their rank ordering in the separate analysis of students with ID only and students with ID and ASD. In addition, estimates of difficulty were stable between the two scales, providing evidence for research question 2, that the thinking skills learning progression could be used for both groups.

3.2.2 Investigation of individual items

In the next step to investigate DIF for students with ASD and ID, the expected score curves (ECCs) were visually investigated. The ECCs were generated via a Rasch partial credit model analysis in Conquest (Adams et al., 2015). The ECCs plot the probability of achieving a criterion along a path of increasing difficulty within an item. The probability of success of students with ID only and students with ASD and ID were plotted together on one graph to visually inspect the differences. Figure 10 illustrates an ECC for Item 6.

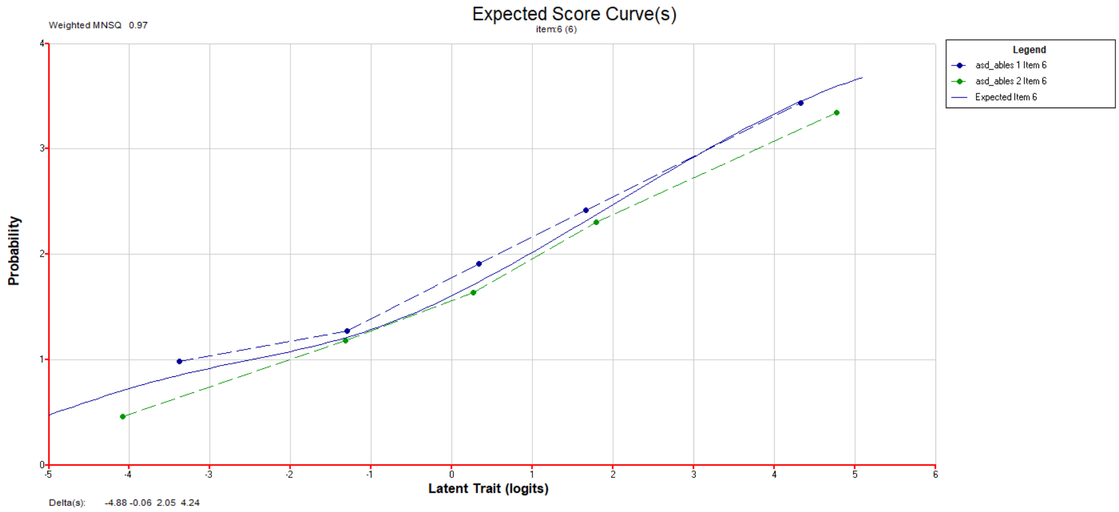

Figure 10. Expected Score Curve for Item 6: Sequencing steps to reach a preferred or relevant outcome in a familiar task.

The probability of students with ID only being successful at different points along the difficulty scale is denoted by the lower dotted line. Students with ASD and ID are denoted by the higher dotted line. This means that students with ID only were slightly more likely to be assigned a criterion at a lower level of difficulty than a student with ID and ASD with the same level of ability.

Figure 11 shows the ECC for Item 14.

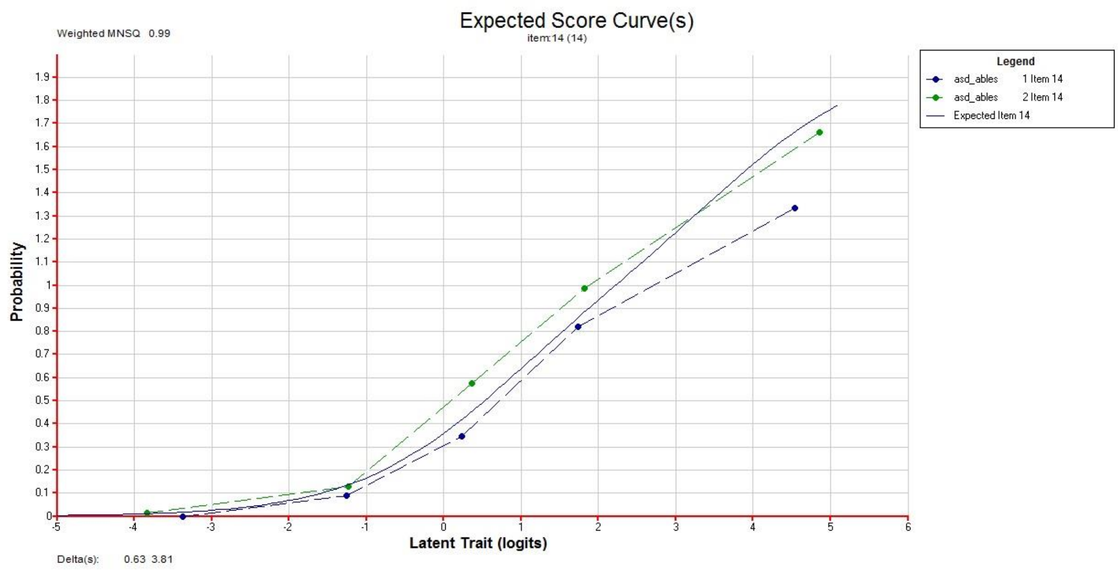

Figure 11. Expected Score Curve for Item 14: Understanding another's viewpoint.

In Figure 11, students with ASD and ID are denoted by the lower dotted line and students with ID are denoted by the higher dotted line. This indicates that students with an ASD diagnosis were consistently slightly more likely to be assigned a criterion at a lower level of difficulty for this item than a student at the same ability level without an ASD diagnosis.

The ECCs demonstrated the same patterns of difference as shown in the bi-variate plots above. However, such differences could be due to different strengths and weaknesses of students in the separate groups and not due to systematic bias towards a particular group. Thus, the result of visual inspection of the bi-variate plots and the ECCs was the flagging of items listed below for Mantel Haenszel (MH) analysis.

- Item 1: Focusing on relevant information
- Item 4: Choosing a preferred or relevant outcome
- Item 5: Dividing a new or unfamiliar task into parts or steps
- Item 6: Sequencing steps to reach a preferred or relevant outcome in a familiar task

- Item 13: Understanding another’s intention
- Item 14: Understanding another’s viewpoint
- Item 15: Developing strategies

The bivariate plots and ECCs indicated that Items 5 and 6 were slightly less difficult for students with ID and ASD and showed minimal difference with Item 1. Conversely, Items 4, 13, 14 and 15 were slightly more difficult for students with ID and ASD. They align with theories describing challenges for students with ASD such as difficulties with flexibility and planning (Happé et al., 2006; O'Hearn et al., 2008; Ziermans et al., 2017) described by Items 4 and 15 and theory of mind (Baron-Cohen, 2010; Frith, 2008) described by Items 13 and 14.

3.2.3 Mantel Haenszel analysis

MH analysis was carried out on the flagged items listed above to check differential step functioning. The analysis on Items 1, 4, 5, 6, 13, 14 and 15 indicated the same pattern of strengths and weaknesses as found in the investigation of the ECCS. Nonetheless, the conclusion was that differences were negligible between the two groups.

For instance, the results of the MH analysis for Item 14 is shown in Table 2.

Table 2

MH Statistical Output for Item 14

Step	MH Statistic	Chi square	df	p	DIF category
Between categories 0 and 1	0.002	0.007	4	0.999	Negligible
Between categories 1 and 2	0.451	1.476	4	0.831	Negligible

The output in Table 2 indicated that the item step between Category 0 (Is moving towards but has not yet achieved these skills/behaviours) and Category 1 (Explains or gives information in a situation where an adult may not understand or have all the same information) was estimated to be 0.002 logits higher in difficulty for students with ID and ASD than for students with ID only with a similar level of ability. Accordingly, the item step between Category 1 and Category 2 (Generalises previous experience and knowledge to explain why individuals have different information or feelings) was estimated to be 0.451 logits higher in difficulty for students with ID and ASD than for students with ID only with a similar level of ability. However, both results were negligible in terms of significance.

3.2.4 Expert review

A final check was carried out in light of the statistical analysis to investigate how items worked for students with ID only and with ID and ASD. A review took place of the flagged items to check for sources of bias. A careful review of the wording of the items and possible sources of construct irrelevant variance resulted in a judgement that the items were not systematically biased against students with ID and students with both ID and ASD.

The results of the visual inspection of the bi-variate plots, the ECCs, MH analysis and expert review indicated no bias or effect on the order of student ability estimates. The results of these steps addressed research question number 2 by providing evidence that the derived thinking skills learning progression applied to students with ID who also have ASD. The inspection of the ECCS and the MH analysis detected differences in students' assessed thinking skills that were consistent with challenges that may be faced by students with ASD but these did not impact on the integrity of the calibrated scale.

4. Discussion

This study described a DIF analysis as part of the design and validation of an assessment. Despite a popular tendency to frame the development of children with ASD as atypical (Frith, 2008), this study added to a research base demonstrating that learning progressions describing the order of emergence of foundational skills such as thinking and literacy (Woods & Pavlovic, 2016) can be used for all students including students with ID only and ID and ASD. More specifically, the order of item difficulty was demonstrated to maintain their rank ordering among students with ID, both with and without ASD. The results of the DIF analysis supported literature arguing that skills such as planning, flexibility, or theory of mind can be challenging for students with ASD (Baron-Cohen, 2010; Frith, 2008; Happé et al., 2006; Hill, 2004; O'Hearn et al., 2008; Ziermans et al., 2017). However, these flagged challenges were not so significant that an assessment designed for students with a range of disabilities could not be used to monitor learning for a subgroup who had both ID and ASD.

An implication of the results of the DIF analysis is a pedagogical shift in thinking to a focus on ability rather than on a label or diagnosis. A tendency exists to focus on deficits of students with an ASD diagnosis (Baron-Cohen, 2017; Begeer et al., 2010; Chen et al., 2016) and diagnosis has increased (APA, 2013). While such work is vital, increased focus on diagnosis can perpetuate the exclusion of students with a disability diagnosis as teachers' perception of their ability to teach a student can change with the addition of a diagnostic label (Slee, 2011). The medical model of disability has perpetuated a deficit approach to students with ASD. This is contested in literature that has argued that such approaches are socially constructed (Oliver, 2013; Milton, 2012). Milton (2012) argued that best practice for students with ASD should continually be questioned by educators through consideration of the impact of their own perspectives on the instruction of students with ASD. This study was one step in supporting teachers by providing a resource to target instruction and by providing evidence that all students, including students with an ASD diagnosis develop in much the same way in thinking skills needed to learn and access the general curriculum. It is hoped that this resource will shift educators' thinking to focus on the teaching and learning of thinking skills to all students, while acknowledging unique patterns of challenges and strengths.

One implication of the developed empirically validated scale to describe the learning of students with ID and ASD in thinking skills is that it provides an opportunity to study outcomes of instruction in thinking skills. For instance, studies could examine how progress in thinking skills affects other academic and social outcomes. The developed assessment could also be the basis for work demonstrating the impact of instruction on the development of thinking skills. Such work would provide empirical evidence to augment the emerging research that is showing the benefits of teaching such skills to all students, including students with disability (Diamond & Lee, 2011; Kistner et al., 2010). It would also build on the argument regarding the malleability of such skills (Au et al., 2014) by providing a way to investigate the value of targeting and teaching such skills to students with disability.

It has been established that all students, including students with ID and ASD should access the general curriculum (UN, 2018). Thinking skills are core to being able to access learning in all aspects of the curriculum. The study described in this paper was a step to fill a gap in research, to describe how thinking skills progress in students with ID and ASD working at a level not described in present general curricula. The assessment developed in this study presented a resource for implementation by systematically operationalising thinking skills and targeting levels of skill not previously addressed. Assessment data from 301 students with an ASD diagnosis with a co-occuring diagnosis of ID were shown to progress along a typical trajectory of thinking skills albeit with distinct patterns of strengths and challenges. Such findings provide evidence that the teaching and learning of thinking skills can be implemented for all students. The defined skills were based on developmental readiness, not on curriculum assumptions of competency.

There are limitations to the study that can be addressed in further work. The sample was constrained to participants in Victoria, Australia and defined in terms of the diagnostic categories set out there. The heterogeneity of students with ID and ASD has been well established. As such, with a greater sample, further investigation into more fine-grained groupings of students could be carried out to strengthen the generalisability of the findings. Moreover, as validation is a never-ending task, work is ongoing to continue to provide evidence for the validity of the thinking skills learning progression. For instance, further work would be beneficial to check the generalisability of the assessment in different contexts and over time.

The outcome of the DIF analysis as well as the development of the assessment involving input from SMEs and empirical analysis was made accessible to teachers through an assessment based on a learning progression of thinking skills. The purpose of the assessment was to provide a resource that teachers could use to assess where a student was working and to know where the student would need to proceed next to progress. Such work is important to add focus to learning and teaching for all students according to level of development in contrast to focusing on a diagnosis or label.

Declarations of interest: none

5. References

Adams, R.J., & Khoo, S. (1996). *Quest: The interactive test analysis system*. Camberwell, Australia: Australian Council for Educational Research.

Adams, R.J., Wu, M.L., & Wilson, M.R. (2015). *ACER ConQuest: Generalised item response modelling software* [Computer software]. Camberwell, Australia: Australian Council for Educational Research. Retrieved from: https://www.acer.org/conquest

American Educational Research Association (AERA). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Psychiatric Association (APA). (2013). *Diagnostic and statistical manual of mental disorders, fifth edition (DSM-5).* Washington, DC: American Psychiatric Association.

Anderson, L.W. & Krathwohl, D.R. (Eds.) (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Longman.

Annevirta, T., & Vauras, M. (2006). Developmental changes of metacognitive skill in elementary school children. *The Journal of Experimental Education, 74*(3), 197-225.

Au, J., Sheehan, E., Tsai, N., Duncan, G.J., Buschkuehl, M., and Jaeggi, S.M. (2014). Improving fluid intelligence with training on working memory: A meta-analysis. Psychonomic Bulletin and Review, 22(2): 366-377. doi: 10.3758/s13423-014-0699-x

Australian Curriculum, Assessment and Reporting Authority (ACARA) (2020). *The Australian Curriculum.* Accessed from: https://www.australiancurriculum.edu.au/

Bae, Y. S., Chiang, H., & Hickson, L. (2015). Mathematical word problem solving ability of children with autism spectrum disorder and their typically developing peers. *Journal of Autism and Developmental Disorders, 45*(7), 2200-2208. doi:10.1007/s10803-015-2387-8

Baillargeon, R., Li, J., Gertner, Y. & Wu, D. (2010) How do infants reason about physical events? In U. Goswami, (Ed.) *Part II Cognitive Development in Early Childhood Wiley Blackwell Handbooks of Developmental Psychology: Wiley-Blackwell Handbook of Childhood Cognitive Development.* (p.p. 11-48) Hoboken, GB: Wiley-Blackwell.

Baron-Cohen, S. (2010). The Empathising-Systemising (E-S) Theory of Autism. In U. Goswami (Ed.) *Part II Cognitive Development in Early Childhood Wiley Blackwell Handbooks of Developmental Psychology: Wiley-Blackwell Handbook of Childhood Cognitive Development.* (p.p. 626-639) Hoboken, GB: Wiley-Blackwell.

Baron-Cohen, S. (2017). Editorial perspective: Neurodiversity – a revolutionary concept for autism and psychiatry. Journal of Child Psychology and Psychiatry 58(6), 744-747. doi: 10.1111/jcpp.12703

Begeer, S., Gevers, C., Clifford, P., Verhoeve, M., Kat, K., Hoddenbach, E., & Boer, F. (2011). Theory of mind training in children with autism: A randomized controlled trial. *Journal of Autism and Developmental Disorders, 41*(8), 997-1006. http://dx.doi.org.ezp.lib.unimelb.edu.au/10.1007/s10803-010-1121-9

Bent, C., Barbaro, J., & Dissanayake, C. (2017). Change in autism diagnoses prior to and following the introduction of DSM-5. *Journal of Autism & Developmental Disorders, 47*(1), 163-171. doi:10.1007/s10803-016-2942-y

Berger, S. E., Chin, B., Basra, S., & Kim, H. (2015). Step by step: A microgenetic study of the development of strategy choice in infancy. *British Journal of Developmental Psychology, 33*(1), 106-122. doi:10.1111/bjdp.12076

Bijvoet-van, d. B., & Hoicka, E. (2014). Individual differences and age-related changes in divergent thinking in toddlers and preschoolers. *Developmental Psychology, 50*(6), 1629-1639. doi: 10.1037/a0036131

Burack, J.A., Russo, N., Flores, H., Iarocci, G., & Zigler, E. (2012). The more you know the less you know, but that's OK: Developments in the developmental approach to intellectual disability. In J.A., Burack, R.M. Hodapp, G. Iarocci, G. & E. Zigler (Eds.), *The Oxford handbook of intellectual disability and development*. (pp. 1-14). New York, NY: Oxford Handbooks Online. doi: 10.1093/oxfordhb/9780195305012.013.0001

Chen, S.F., Chien, Y.L., Wu, C.T., Shang, C.Y., Wu, Y.Y., & Gau, S.S. (2016). Deficits in executive functions among youths with autism spectrum disorders: an age-stratified analysis. *Psychological Medicin 46*(8), 1625-1638. doi:10.1017/S003329171502238

Clark, M., Barbaro, J., & Dissanayake, C. (2017). Continuity and Change in Cognition and Autism Severity from Toddlerhood to School Age. *Journal of Autism & Developmental Disorders, 47*(2), 328-339. doi:10.1007/s10803-016-2954-7

Coles-Janess, B., & Griffin, P. (2009). Mapping transitions in interpersonal learning for students with additional needs. Australasian Journal of Special Education, 33(2), 141-150. doi: 10.1375/ajse.33.2.141

Colmar, S.; Maxwell, A. & Miller, L. (2006) Assessing intellectual disability in children: are IQ measures sufficient, or even necessary? *Australian Journal of Guidance and Counselling, 16*(2) 177-188. https://search-informit-com-au.ezp.lib.unimelb.edu.au/fullText;dn=156978

Darling-Hammond, L. (2012) Policy frameworks for new assessments. In P. Griffin, B. McGaw, & E. Care (Eds.), Assessment and Teaching of 21st Century Skills (pp. 301-339) Dordrecht: Springer. https://doi.org/10.1007/978-94-007-2324-5_6

Dewey, J. (1910). *How we think*. [electronic resource]. Boston: D.C. Heath & Co., 1910.

Dewey, J. (1916). *Democracy and Education: an introduction to the philosophy of education. [electronic resource]*. Project Gutenberg, 1997.

Diamond, A., & Lee, K. (2011). Interventions shown to aid executive function development in children 4–12 years old. *Science 333*(6045), 959–963. doi: 10.1126/science.1204529

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, N.J.: L. Erlbaum Associates, 2000. Retrieved from https://ebookcentral.proquest.com

Flavell, J. H., Green, F. & Flavell, E. (1995). *Young children's knowledge about thinking*, Chicago, Ill.: University of Chicago Press, c1995.

Frith, U. (2008). *Autism: a very short introduction*. [electronic resource] Oxford: Oxford University Press, 2008.

Garcia, E. (2015). Lorenzo P. v. Riles? Should the "Larry P." Prohibitions Be Extended to English Language Learners?: Considering Public Policy & IQ Testing in Schools. *Multicultural Education*, *22*(2), 2-7. ERIC number: EJ1065426

Gersten, R., Harris, K. R. and Mason, L. H. (2017). Beyond reading: The less addressed aspects of research in learning disabilities-Introduction to the special issue. *Learning Disabilities Research & Practice*, *32*(3) 137-139. doi:10.1111/ldrp.12145

Glaser, R. (1997). Instructional technology and the measurement of learning outcomes: Some questions (1963). In J. M. Notterman, J. M. Notterman (Eds.), *The evolution of psychology: Fifty years of the American Psychologist* (pp. 337-342). Washington, DC, US: American Psychological Association. doi:10.1037/10254-024

Griffin, P. (1993). *Profiles: Assumptions and procedures in their development.* Melbourne: Assessment Research Centre, RMIT.

Griffin, P. (2007). The comfort of competence and the uncertainty of assessment. *Studies in Educational Evaluation, 33*, 87–99.

Griffin, P., Gillis, S., & Calvitto, L. (2004). Connecting competence and quality: Scored assessment in Year 12 VET. Retrieved from https://www.researchgate.net/publication/238777724_Connecting_Competence_and_Quality_Scored_Assessment_in_Year_12_VET

Guilford, J. P. (1954). The method of paired comparisons. In J. P. Guilford (Ed.), *Psychometric methods* (pp. 154-177). Toronto, Canada: McGraw-Hill.

Happé, F., Booth, R., Charlton, R., & Hughes, C. (2006). Executive function deficits in autism spectrum disorders and attention-deficit/hyperactivity disorder: Examining profiles across domains and ages. *Brain And Cognition, 61(Special Issue: The Cognitive Neuroscience of Asperger Syndrome)*, 25-39.

Hill, E.L. (2004). Evaluating the theory of EF deficits in autism. *Developmental Review, 24*(2004), 189-233. doi: 10.1016/j.dr.2004.01.001

Hill M., W., Marquez, B., Yeaton, P., Pennefather, J., Forness, S.R., & Vincent, C.G. (2015). Teacher judgment in assessing students' social behavior within a response-to-intervention framework: Using what teachers know. *Education and Treatment of Children 38*(3), 363-382. doi:10.1353/etc.2015.0019

Hock, M. F., Bulgren, J. A. & Brasseur-Hock, I. F. (2017), The Strategic Instruction Model: The less addressed aspects of effective instruction for high school students with learning disabilities. *Learning Disabilities Research & Practice, 32*(3) 166-179. doi:10.1111/ldrp.12139

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* London: Routledge and Kegan Paul.

Kane, M. (2016) Validation strategies delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. R. Raymond, & T. M. Haladyna, (Eds.) *Handbook of test development*. (pp. 64 – 80). New York, NY: Taylor and Francis.

Kistner, S., Rakoczy, K., Otto, B., Dignath-van Ewijk, C., Buttner, G., & Klieme, E. (2010). Promotion of self-regulated learning in classrooms: investigating frequency, quality, and consequences for student performance. *Metacognition and Learning 5*(2010)157-171. Doi: 10.1007/s11409-010-9055-3

Koslowski, B. & Masnick, A. (2010). Causal reasoning and explanation. In U. Goswami, (Ed.) *Part II cognitive development in early childhood Wiley Blackwell handbooks of developmental psychology: Wiley-Blackwell handbook of childhood cognitive development (2).* (pp. 377-398) Hoboken, NJ: Wiley-Blackwell.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149. doi:10.1007/BF02296272

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement 3rd ed*., (pp. 13-103). New York, NY: MacMillan Publishing.

Milton, D. E. M. (2012). On the ontological status of autism: the 'double empathy problem'. *Disability & Society 27*(6), 883-887. DOI: 10.1080/09687599.2012.710008

O'Hearn, K., Asato, M., Ordaz, S., & Luna, B. (2008). *Neurodevelopment and executive function in autism. Development and Psychopathology, 20*(4), 1103-1132. doi:10.1017/S0954579408000527

Oliver, M. (2013) The social model of disability: thirty years on, *Disability & Society, 28*(7), 1024-1026, DOI: 10.1080/09687599.2013.818773

Organization for Economic Co-operation and Development (OECD) (2018). The future of education and skills Education 2030. Accessed from https://www.oecd.org/education/2030/E2030%20Position%20Paper%20(05.04.2018).pdf

Osterlind, S.J. & Everson, H.T. (2011). *Differential Item Functioning*. SAGE Publications, Inc. https://dx.doi.org/10.4135/9781412993913

Parmenter, T. R. (2011). What is intellectual disability? How is it assessed and classified?. *International Journal of Disability, Development and Education, 58*(3), 303-319. doi:10.1080/1034912X.2011.598675

Penfield, R. D., Myers, N.D., Wolfe, E.W. (2008). Methods for assessing item, step, and threshold Invariance in polytomous items following the partial credit model. *Educational and Psychological Measurement 68*(5): 717-733. doi: 10.1177/0013164407312602

Piaget, J. (1954). *The construction of reality in the child*. (M. Cook, Trans.). New York, NY: Basic Books.

Rasch, G. (1960) *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Neilson & Lydiche.

Roberts, E., & Griffin, P. (2009). Profiling transitions in emotional development for students with additional learning needs. Australasian Journal of Special Education, 33(2), 151-161. doi: 10.1375/ajse.33.2.151

Rueda, M. R., Checa, P., & Cómbita, L. M. (2012). Enhanced efficiency of the executive attention network after training in preschool children: Immediate changes and effects after two months. Developmental Cognitive Neuroscience, 2(Supplement 1), S192-S204. doi:10.1016/j.dcn.2011.09.004

Rychen, D.S. & Salganik, L.H. (eds.) (2003). *Key Competencies for a Successful Life and Well-Functioning Society.* Hogrefe & Huber, Göttingen.

Siegler, R. S., DeLoache, J. S., Eisenberg, N., Saffran, J., & Leaper, C. (2014). *How children develop*. New York, NY: Worth Publishers, a Macmillan Higher Education Company.

Slee, R. (2011). *The irregular school: Exclusion, schooling, and inclusive education*. London; New York: Routledge.

Spencer, V. G., Evmenova, A. S., Boon, R. T., & Hayes-Harris, L. (2014). Review of research-based interventions for students with autism spectrum disorders in content area instruction: Implications and considerations for classroom practice. *Education and Training in Autism and Developmental Disabilities, 49*(3), 331-353.

Taylor, J.C., Rizzo, K.L., Hwang, J. & Hill, D. (2019). A review of research on science instruction for students with autism spectrum disorder. School Science and Mathematics. 2020(120), 116-125. Doi: 10.1111/ssm.12388.

United Nations (UN) (2018). *Sustainable Development Goals 4 Quality Education*. Retrieved from https://www.un.org/sustainabledevelopment/education/

Victoria State Government. (2016). *Review of the Program for Students with Disabilities.* Retrieved from http://www.education.vic.gov.au/Documents/about/department/PSD-Review-Report.pdf

Victorian Department of Education and Training (2018) *The Abilities Based Learning and Education Support (ABLES)*. Retrieved from: http://www.education.vic.gov.au/school/teachers/learningneeds/Pages/ables.aspx

Vygotsky, L.S. (1980). *Mind in society: The development of higher mental processes*. (Eds. M. Cole, V., John Steiner, S., Scribner, & E. Souberman). Cambridge, MA: Harvard University Press.

Wellman, H (2010). Developing a theory of mind. In U. Goswami (Ed.) *Part II Cognitive Development in Early Childhood Wiley Blackwell Handbooks of Developmental Psychology: Wiley-Blackwell Handbook of Childhood Cognitive Development.* (p.p. 258-284). Hoboken, GB: Wiley-Blackwell.

Wilson, M. (2005). *Constructing measures: an item response modeling approach*. Mahwah: Lawrence Erlbaum Associates.

Wise, L.L. & Plake, B.S. (2016) Test Design and Development Following the Standards for Educational and Psychological Testing. In S. Lane, M.R. Raymond, & T.M. Haladyna, (Eds.) *Handbook of Test Development*. [electronic resource]. (pp. 19-39) Florence: Taylor and Francis

Wolfe, E.W., & Smith, E.V., Jr. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I – Instrument development tools. In E.V. Smith, Jr. & R. M. Smith (Eds.). *Rasch measurement: Advanced and specialized applications* (pp. 202 – 242). Maple Grove, MN: JAM Press.

Wolfe, E.W., & Smith, E.V., Jr. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II – Validation activities. In E.V. Smith, Jr. & R. M. Smith (Eds.). *Rasch measurement: Advanced and specialized applications* (pp. 243 – 290). Maple Grove, MN: JAM Press.

Woods, K., & Griffin, P. (2013). Judgment-based performance measures of literacy for students with additional needs: Seeing students through the eyes of experienced special education teachers. Assessment in Education: Principles, Policy & Practice, 20(3), 325 – 348. doi: 10.1080/0969594X.2012.734777

Woods, K. & Pavlovic, M. (2016). Using differential item functioning to validate a judgement-based assessment of emergent literacy for students with autism spectrum disorder. Presented at Australian Association for Research in Education Conference. Melbourne, 2016.

Wright, B. D. & Stone, M.H. (1979). *Best test design.* Chicago: Mesa Press.

Wu, M. & Adams, R. (2007). Applying the Rasch model to psycho-social measurement: A practical approach. Educational Measurement Solutions, Melbourne.

Wu, M., Tam, H. P., & Jen, T. (2016). *Educational measurement for applied researchers: theory into practice*. Singapore: Springer.

Zelazo, P.D. & Müller, U. (2010) Executive Function in Typical and Atypical Development. In U. Goswami (Ed.) *Part II Cognitive Development in Early Childhood Wiley Blackwell Handbooks of Developmental Psychology: Wiley-Blackwell Handbook of Childhood Cognitive Development (2).* (p.p. 574-603) Hoboken, GB: Wiley-Blackwell.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 592–611). Hillsdale, NJ: Erlbaum.

Ziermans, T., Swaab, H., Stockmann, A., de Bruin, E., & van Rijn, S. (2017). Formal thought disorder and executive functioning in children and adolescents with autism spectrum disorder: Old leads and new avenues. Journal of Autism and Developmental Disorders, 47(6), 1756-1768. http://dx.doi.org.ezp.lib.unimelb.edu.au/10.1007/s10803-017-3104-6

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*(3), 185-197. doi:10.2307/1165031

Zwick, R. (2012). *A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement*. Educational Testing Service Research Report RR-12-08.