



Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Hattie, JAC;Brown, GTL;Earl Irving, S

Title:

An Analysis of an Assessment Tool for 5-year Old Students Entering Elementary School:
The School Entry Assessment Kit

Date:

2015-05-01

Citation:

Hattie, J. A. C., Brown, G. T. L. & Earl Irving, S. (2015). An Analysis of an Assessment Tool for 5-year Old Students Entering Elementary School: The School Entry Assessment Kit. New Zealand Journal of Educational Studies, 50 (1), pp.87-105. <https://doi.org/10.1007/s40841-015-0001-4>.

Persistent Link:

<https://hdl.handle.net/11343/283267>

An analysis of an assessment tool for 5-year old students entering elementary school:
The School Entry Assessment Kit

An analysis of an assessment tool for 5-year old students entering elementary school:
The School Entry Assessment Kit

John A. C. Hattie

Gavin T. L. Brown

S. Earl Irving

University of Melbourne

University of Auckland

Running Head: School Entry Assessment

Address Correspondence to

John Hattie

MERI, University of Melbourne

jhattie@unimelb.edu.au

Abstract

In the Western world, most five year old children begin their formal schooling, and do so with diverse proficiencies. A common aim is for teachers to quickly evaluate their proficiencies to identify strengths

and gaps, and then begin the teaching of reading and number. This article reports two studies into the item and structural characteristics of a well used instrument administered in the first 6-8 weeks of the students' school experience. The School Entry Assessment Kit consists of *Concepts about Print* (CAP) measure of early literacy, *Tell Me* procedures for oral language and story retelling, and *Checkout* game for early number knowledge and ability. The first study reviews the properties of these tests, and the second study relates the performance on the SEA with performance 1 and 5 years later. Implications for measuring progress, reviewing the tests, enhancing reporting, and reducing administration time are discussed.

An analysis of an assessment tool for 5-year old student entering elementary school:

The School Entry Assessment Kit

Starting school is a major milestone for children, and it is well documented that the students start school with very different prior experiences, understanding and knowledge, and remarkable variations in language and numeracy development. This study looks at one system and how it enables teachers of first year children to understand this variation so as to best prepare lessons to take into account what the children bring into their classrooms.

The vast majority of students in New Zealand start school upon their 5th birthday and entry to school is permitted throughout the school year. New Zealand primary schools generally provide 'new entrant' classes for 5 year olds where students are socialised into school procedures and their skills in literacy, numeracy, and oral language are ascertained. Most students would spend about 10 weeks (one term) in such reception classes before being integrated into Year 1. During their first six-eight weeks in schools, it is common for students to be tested on the School Entry Assessment (SEA) kit (Ministry of Education, 1997a, 1997b). The SEA consists of Marie Clay's *Concepts about Print* (CAP) measure of early literacy, McNaughton's (1995) *Tell Me* procedures for oral language, story retelling, and Young-Loveridge's (1997) *Checkout* procedure for early number knowledge and ability.

The SEA is a low-stakes measure intended to inform teacher and school practice. Some schools voluntarily report the scores to the Ministry of Education to be used in system monitoring research (Gilmore, 2004; Philips, 2000; Dewar & Telford, 2003). Dewar and Telford (2003) reported that over half of primary schools reported using the SEA in 2001, with about a third reporting their data to the Ministry. They found that the majority of teachers who used the SEA (86%) always used the CAP, while just under half (46%) always used the *Tell Me* or *Checkout* instrument. The latter result was attributed to the time and difficulty experienced by the teachers in administering these two instruments. Two-thirds of teachers in 2000 indicated that the complete SEA needed revision and updating.

There have been a number of evaluations of the SEA. Anderson, Lindsey, Schulz, Monseur, and Meiers (2001) evaluated the technical and methodological aspects of the SEA and reported issues with the underlying theoretical framework of the SEA, insufficient support to teachers for the effective use of the diagnostic data, the lack of linkage of the SEA to later measures of achievement, and some problems with the measurement characteristics of *Tell Me*. They reported that about 85% of students were assessed within the first eight weeks of enrolment and that later administration had a small but positive enhancement of scores only on the Concepts about Print instrument. Reliability estimates from the 2000

data computed by Anderson, et al. (2001) were generally very high (*CAP* alpha = 0.85; *Tell Me* alpha = 0.90; *Checkout* alpha = 0.85). Inter-test correlations were reported as moderate, ranging from 0.39 to 0.77 (Anderson, et al., 2001). *Tell Me* was criticised especially because of the confounding effect of students selecting a different book and for the subjective scoring of the comprehension score; a conclusion also reached by the Education Review Office (ERO, 2000).

The goal of this article is to report two psychometric studies into the item and structural characteristics of the instruments, which lead then to specific recommendations for the enhancement of these assessments. In Study 1, factor and mean analyses of each instrument and of the battery of instruments are reported, along with a two-parameter IRT based item analyses of the CAP. The differential performance of students by age, sex, and ethnicity on the three instruments is reported. Study 2, examined the predictive validity of SEA scores by comparing performance at age 5 with performance on other measures of literacy and numeracy learning at ages 6 and 10, respectively.

Instruments

The *Concepts about Print* task was developed by Dame Marie Clay as part of a comprehensive survey of children's progress in the first two years of school, and aims to "check children's progress in reading and writing" and has "proved to be a sensitive indicator of one group of behaviours that supports reading acquisition" (Ministry of Education, 1997b, p. 9). It involves a series of 24 questions, each scored dichotomously, relating to a book about "Sand" or "Stones" focused on assessing children's ability to order words, recognise grammar of reading, and placement of text.

Tell Me involves the child sharing a story (out of six possible books) with the teacher and/or a friend, and being assessed as to their ability to orally retell the story that was read aloud. A retelling is used because children's retelling skills are linked to their understanding and skill in other areas of language and literacy. Retelling skills are associated with comprehension, reading for meaning, and producing stories (McNaughton, 1995). The six scored items relate to the comprehension of the story, the complexity of the sentences used, students' use of vocabulary, how they organise the retelling, the context included in the retelling, and the degree of description and expression in their retelling.

Checkout assesses a child's numeracy skills and understandings through their recognition of numerals, their recognition of patterns, their ability to form sets of numbers up to 12, their knowledge of the number sequence, and their ability to mentally add and subtract objects. This is achieved through a shopping game which provides a familiar context to assess the child's skill in numeracy. The five items have multiple score points (i.e., 6, 3, 6, 7, 5) for a total maximum score of 27.

The Study 1 data used in these psychometric studies comes from a dataset of scores submitted to the Ministry of Education from the voluntary submission by schools of their data between 1997 and 2004. The set contains 53,140 students (48% female, 52% male) who completed at least one of the SEA instruments. There were between 6,000 and 10,000 students in each year. The ethnic mix of the sample is very representative of the New Zealand population: NZ European or Pakeha (61.4% compared to population percentage of 54.7%), Maori (22.2% compared to 24.4%), Pasifika (10.4% compared to 10.1%): Asian (4.8% compared to 8.7%), and Other (1.7% compared to 2.1%). Anderson et al. (2001) reported that the school sample between 1998 and 2000 was biased to only a small extent with the year 2000 sample coming from higher socio-economic status schools, the 1999 sample coming from more urban schools, and consistently the samples represented larger schools. In contrast, they found that the samples were not meaningfully biased for student sex or ethnicity. Thus, the sample is large enough to provide generalisations about the performance of ethnic groups and sexes, as well as providing sufficient data to determine the psychometric characteristics of the items and instruments.

Study 2 makes use of data from more than 2000 students from five schools in New Zealand, participating in a longitudinal design over a 10+ year period commencing with the SEA at age 5 until leaving secondary school. At the time of this analysis, data were available from three time-points: age 5 SEA, age 6 Year Net, and age 10 asTTle reading and mathematics tests (Assessment Tests for Teaching and Learning). Since the data were from a cluster of schools in one region, the data are indicative and not claimed to be generalisable.

The data set for each instrument is analysed in turn, before reporting results across all three instruments. Maximum likelihood exploratory factor analysis with oblique (correlated) rotation is used to determine the factor characteristics within and across all three instruments. Given the sample sizes, it was decided to use effect sizes as well as tests of statistical significance to identify the practically significant differences for student demographic characteristics (i.e., sex, ethnicity, home language). Hattie (2009, 2012) has analysed over 150 influences on achievement, based on 900+ meta-analyses, and found that an effect-size of 0.40 is average. Any effect-size less than 0.20 can be considered small, and greater than 0.60 considered large. Two-parameter item response analysis, using BILOG, is used to determine the difficulty and discrimination of each item within the Concepts about Print instrument (Hambleton, Swaminthan & Rogers, 1991). To determine the predictive validity of the SEA scores, Pearson correlations are reported between SEA scores and measures of academic achievement reported one and four years later for two subsets of students.

Results

Study One: Psychometric study. The first validity study examines the relationship of the scales scores for the three SEA instruments. As would be expected, a maximum likelihood factor analysis of all components of the SEA indicated that there were three clear factors (Table 1). The first factor related to *Tell Me*, the second to *Checkout*, and the third to *Concepts About Print*. The correlations between the three factors indicated that each test was contributing significant unique variance and, thus, has the potential for providing discriminating and unique information about the child's proficiencies in the first weeks of school.

Concepts about Print. Between 1998 and 2004 there was a very small increase in the total score from about 9.50 out of 24 to around 10.0. Table 2 presents the means and spread (sd) for each item, and the contribution each item makes to the total score (Item-Total r). There were three clear factors in the CAP, confirming the intended structure (Table 2). The first related to **word** ordering (reordering and sequencing) knowledge, the second to concepts about **language** and the grammar of reading, and the third to the correct **orientation** of texts. In the subsequent analyses students' total scores on these three factors were used. The correlations between the factors were such that it is reasonable to conclude that these three factors contributed meaningful information (language to word $r = 0.57$, language to orientation $r = -0.06$, and word to orientation $r = 0.29$). The CAP can provide useful diagnostic feedback not only at the total score level, but also at the level of these three factors.

By creating a mean item factor score, it was found that the students scored highest on text orientation ($M=0.79$), then language about reading ($M=0.36$), and lowest on word ordering ($M=0.05$). Thus, most students start schools with a higher level of knowledge about the way texts are oriented, while very few have any sense of word or letter sequencing and reordering. There were no meaningful differences between boys and girls on either the total CAP score (effect-size = 0.16 in favour of girls), nor on the three factor scores (effect-sizes: word order = 0.09, language = 0.16, and text orientation = 0.11). It is worth noting however, that by the time students get to Year 5, the difference in literacy reading comprehension between sexes has almost trebled (e.g., an effect size of 0.43 was reported in Leeson, Brown, Hattie, et al., 2005).

There were major differences between those who came from homes where a language other than English was dominant and those from English-dominant homes. The effect-size was 0.59 for the total score, and 0.16, 0.63, and 0.39 for the three factors. This difference was also found to be of similar scale at the end of Year 6 (i.e., $d=0.44$ in Leeson, Brown, Hattie, et al., 2005). Those who had a language other

than English at home had much more difficulty with the grammar of words, less so with the placement of words, and all students regardless of dominant home language found the word ordering difficult.

The database supplied the specific language spoken at home and major differences depending on the language were found, though the sample sizes were very small for some languages (Table 3). The final column indicates the effect-size compared to the largest group – those with English at home. For example, the effect-size between English at home and Maori at home was about 0.5, while between English at home and Pasifika languages at home it was closer to 0.8. It is clear that those from Pasifika homes have a greater disparity in their reading preparedness when they enter school.

There were ethnicity differences, with NZ European/Pakeha students outscoring all other groups (Table 4). The effect size differences show that Maori (-0.60 behind) and Pasifika (-0.84 behind) student groups are substantially behind NZ European/Pakeha students at the start of schooling, while Asian groups are much closer to NZ European/Pakeha children when they begin school (-0.27 to -0.32 behind). An effect size of 0.60 is roughly equivalent to two years' learning (Hattie, 2009). Thus, Maori students start school about two years behind Pakeha when they commence school, and Pasifika students are about 3 years behind.

An item analysis was completed to better understand where students had more difficulties (Table 5). A two-parameter item response model was used to estimate the item difficulties and discriminations. The difficulty parameters (b) are expressed on a logit scale where a doubling of logits represents a doubling of difficulty. The measures of discrimination (a) indicate how successful the items were in finding a difference between the more and less proficient students. The discriminations were quite varied, ranging from 0.22 (item hardly discriminates between the more and less proficient students) to 4.17, with an average value of 1.22.

The plots of the item characteristic curves (ICC) showed that most items performed in the desired manner, but there were some items that merited re-working or removal, especially Items 10, 12, 13, 14, 17, 18, and 20 as they were lacking in discrimination (most of these items relate to Word ordering). These items do not seem to discriminate very well between the lesser and more proficient students and indeed could provide misleading information about students' performance. In contrast, there is significant merit, especially should a short form be required, in retaining items with high discrimination values (i.e., 1 to 7, 9, 19, 22, and 23) as they have significant power to discriminate among 5 year olds; although this would then not include any Word ordering items.

The total test characteristic curve shows that the CAP has very good discrimination power across most levels of reading proficiency, and not surprisingly dips in information for children who display high levels of skill (Figure 1).

In summary, the CAP contains three moderately inter-correlated factors and provides most powerful information about students who are weak to moderate in those skills upon entry to New Zealand primary schools. Most items provided valuable information, although there were some items that were either redundant or uninformative. It may be that two instruments are warranted. The first would be a revised version of the current CAP that keeps the easy, high discriminating, non-redundant items as a measure of fundamental *Concepts about Print* which should be administered only to students who clearly would struggle to answer Item 1 or 2. The second instrument should be designed to clearly identify students who are ready to embark on more challenging reading tasks; this instrument should be more challenging in order to provide diagnostic details about enhanced teaching for these more advanced learners.

Tell Me. Children are offered a choice of six books which are read aloud as the basis of the retelling. It is argued that the choosing of books is important to ensure optimal engagement by the children. The most popular book is Bernard O'Brien's Tooth (39%), then Too Much Cake (16%), Naming the baby (11%), Terri and the Rocket (7.4%), Great times with Great Granddad (5.4%), and The Daisy Chain (2.7%). Similarly to Anderson et al. (2001), this analysis found that there were major differences in the choice of books by boys and girls (Figure 6). Boys preferred Terri and the Rocket (71% of those who choose this book were boys, although boys choose Bernard's Tooth the most often), while girls preferred The Daisy Chain (74% to 26%), and Naming the Baby (67% to 33%).

It has been claimed by Gilmore (2004) that "from a psychometric perspective *Tell Me* should be based on one story only. The only way to determine the different difficulty levels of each story accurately would be to assign stories randomly to the students." This is not correct. Nor is it correct, as claimed by Anderson et al. (2001, p. 19) that "to get proper estimates of the effect of the choice of book it is necessary to have a random distribution of books to students." Every story (as can every item in a test, every reading passage, every sub-section) can be differentially weighted using for example IRT or regression methods to allow for differences across books. The two-parameter item response model, because it estimates individual item discrimination indices, easily handles the differential difficulty of a story, and such statistical procedures can be built into a software resource such that score calculation can be made taking into account the student's choice of book in *Tell Me*. There is a need for a careful

equating study of *Tell Me* if it is educationally important to allow students to choose books, then this can be readily modelled in any test model using item response methods (and thus can resolve Gilmore's criticism). The psychometric model underlying the assessment process should model the desired behaviour (i.e., student's choose the story) rather than the other way around. It is emphasised, however, that these IRT adjustments to take into account the choice of book are critical. As can be seen in Figure 2, there are major differences in Comprehension scores from the easiest to hardest story (effect-size between easiest and hardest ≈ 0.60), while across all others there is closer comparisons). Students find it hardest to comprehend *The Daisy Chain* and easiest to comprehend *Terri and the Rocket*. Across the other dimensions, *Bernard O'Briens' Tooth* has easier attributes in Vocabulary, Organisation and Description.

From a maximum-likelihood factor analysis, there is only one underlying factor (which could be described as 'early reading skills'), explaining 63% of the total variance, and all but one test loads highly on this first factor: Organisation 0.87, Vocabulary 0.84, Sentences 0.84, Content 0.80, and Comprehension somewhat lower at 0.52. This *hints* at a second factor separating the proficiencies of Comprehension from the others, and more items relating to Comprehension may be worth considering.

There are two stories on which there were no statistically significant differences between the means for boys and girls (*Great times with Granddad*, and *Terri and the Rocket*, even though the latter was chosen by boys much more often), but there were statistically significant differences in the other stories, in all cases favouring girls. The effect-sizes, however, are relatively small (*Granddad* $d=0.03$, *Terri* $d=0.05$, *Bernard* $d=0.10$, *Cake* $d=0.16$, *Daisy* $d=0.20$, *Baby* $d=0.27$). It would be worthwhile to closely investigate "Naming the baby" (a popular choice with girls) to ascertain why this story appeared to have such a consistent bias in favour of girls. While there were statistically significant differences for the six *Tell Me* component scores across ethnicities, the pattern was similar across them all. The NZ European/Pakeha and Maori scores were closest, while the Pasifika students scored systematically the lowest (Figure 3).

In sum, the *Tell Me* materials produced story effects which cannot be adequately addressed without the use of computerised IRT scoring to take into account this choice of book as a major source of score variance. Such an approach would allow for the motivational benefit of allowing students free choice in the selection of reading material while providing comparable scores across stories. If teachers wish to minimise discrepancies relating to gender and ethnicity, then they should choose *The Daisy Chain* or *Great times with Granddad*; then *Too much Cake* and *Naming the baby*; but the stories with the greatest sources of variability are *Terri and the Rocket* and *Bernard O'Brien's Tooth*.

Checkout. Check Out assesses a child's numeracy skills and understandings through their recognition of numerals, their recognition of patterns, their ability to form sets of numbers up to 12, their knowledge of the number sequence, and their ability to mentally add and subtract objects. This is achieved through a shopping game which provides a familiar context to assess the child's skill in numeracy. The distribution of scores for Check Out indicates a near normal distribution across the sample of students. Students had higher means for Sets (5.39, $d=1.79$), Numerals (4.80, $sd=2.06$), Sequences (3.98, $sd=2.20$), Patterns (2.90, $sd=1.23$) then operations (1.85, $sd=1.70$). A maximum-likelihood factor analysis indicated that there was only one factor (basic numeracy): the loadings were Numerals (0.84), Sets (0.82), Sequences (0.77), Patterns (0.66), and Operations (0.55).

There were very small differences between boys and girls across the five sub-scores with effect-sizes in favour of girls being in the range of 0.02 to 0.14. Across all ethnic groups the same pattern of ability existed as in the other two tests (Table 6); that is, Pakeha and Asian, then Maori, then Pasifika (although Cook Islanders scored similar to Pakeha, $d=0.09$ compared with Samoan $d=0.41$ and Tongan $d=0.84$).

Study Two: The predictive value of the SEA. In New Zealand schools, teachers regularly administer the Six Year Net on or about children's sixth birthday. This instrument, developed by Marie Clay, includes the Concepts about Print along with other measures of early literacy. The net involves asking students to carry out specific tasks associated with identifying letters, understanding print concepts, reading text, recognising words, writing vocabulary, and hearing and recording sounds in words. Many schools use this instrument to identify children requiring additional one-on-one reading recovery instruction. The correlations between CAP scores at entry to school and the Six Year Net were very high (r ranging from 0.75 to 0.99), a not unexpected result in that the CAP is also part of the Six Year Net (Table 7). The strongest other correlates of literacy performance at age 6 were the oral language *Tell Me* Comprehension (r ranging from 0.48 to 0.59) and the Numerals score from *Checkout* (r ranging from 0.60 to 0.69). These correlations for two measures taken one year apart accounted for between 25% and 80% of variance.

From such strong correlations, however, it seems that those who started school with better literacy concepts remained so, and those not advantaged remained disadvantaged after the first year of schooling. Such "excellence" in prediction should be a major concern for teachers, and at a minimum more attention should be placed on either ensuring more children start school with the skills underlying the CAP (and oral language and numerals), and/or that teachers are more effective in teaching these skills as early as possible during their first year of schooling.

The next analyses in this longitudinal study related the SEA to Progressive Achievement Tests (PAT) and Assessment Tools for Teaching and Learning (asTTle) scores another 4 years later in Year 5 (Table 8). The PATs are standardised, multiple-choice, norm-referenced measures of general ability in listening comprehension, reading comprehension, reading vocabulary, and mathematics. asTTle provides measures of reading comprehension and mathematics drawn from a bank of nationally-referenced, IRT-calibrated, teaching/curriculum-aligned items. Appropriate forms were administered to students in Year 5 and correlations were found for these tests and their SEA scores obtained five years earlier. The average correlation of the SEA scales with the various PATs was 0.39, with noticeably larger correlations found between the PAT Reading Comprehension and Vocabulary tests and four of the oral *Tell Me* scores (r ranged 0.65 to 0.79). Interestingly, the Sequences scale from *Checkout* were moderately correlated with all four of the PAT tests (r ranged from 0.52 to 0.58), while the Sets scale correlated moderately with the PAT Mathematics and Listening Comprehension tests ($r = 0.62, 0.51$, respectively), and the Checkout total had similar correlations with the Reading Vocabulary and Mathematics tests ($r = 0.52$ in both cases). The three SEA tests were good predictors of later reading vocabulary (mean $r=0.49$), and somewhat less powerful at predicting comprehension (mean $r=0.34$).

In contrast, the correlations between SEA and the asTTle reading and mathematics tests ranged from as low as 0.03 to only 0.44, with an average of 0.28. The *Concepts about Print* had the strongest correlations with reading and mathematics ($r = 0.40, 0.43$ respectively), five of the *Tell Me* scales had an average correlation of 0.38 with the asTTle reading test, and two of the *Checkout* scores and the *Checkout* total had correlations with the asTTle mathematics test ranging between 0.41 and 0.44. Although the pattern of correlations is appropriate (oral language correlates more with reading than mathematics; numeracy correlates with more with mathematics than reading), the strongest correlations only accounted for only 16% to 20% of variance.

The reasons for the higher correlations in the PAT is probably related to these tests measuring underlying abilities (e.g., reading vocabulary is one of the better predictors of IQ scores), whereas the asTTle tests are more aligned with the achievement expected from the NZ curriculum (it is noted that the PAT has since been revised to make them more curriculum oriented). It is harder to change underlying ability than it is to add value in achievement – and the lower correlations should somewhat ameliorate the criticism above that a fundamental role of teachers is to “mix-up” the correlations – and thus allow students who put in effort, who are engaged in schooling, who develop appropriate achievement strategies to then show larger achievement gains. It seems inappropriate to use ability related tests to assess achievement of students, or the proficiency of teachers to teach the curricula.

Table 5 reported the effect sizes between NZ European/Pakeha children at age five and those of other ethnic groups. Five years later, the asTTle reading tests had smaller effect sizes: the gap for Pasifika students (0.84 to 0.40), whereas for Maori the gap reduced less so from 0.60 to 0.49. This provided evidence that, notwithstanding the strong correlations to the PATs, schools can reduce the achievement gaps for Pasifika students and to a lesser extent for Maori students (52% vs. 18% reduction in effect sizes respectively).

To check whether there are different gains for differing achievement levels of students, the students were divided into three groups (about $N=130$ in each) on the three SEA tests and then the mean and standard deviation was estimated one year later (on the 6 Year Net). Table 9 shows that the students in the top two-thirds of performance on the SEA have the highest 6 Year Net scores – and that there are very few differences in the means for these top two groups. Those students in the lowest third of the SEA stayed behind when evaluated a year later on the Six Year Net, and also were the one's falling behind 4-6 years later (as assessed by the PAT and asTTle). The children above the bottom third have variable performance later on – that is, they have sufficient knowledge about reading and mathematics such that teaching, diligence, work ethic, and so on are probably additional important predictors of success than only the ability when they began school.

Perhaps this also suggests that no matter what teachers learn from the SEA, their subsequent use of this information is not making a major difference to those at the bottom of the achievement distribution! The same data can be plotted to show the gains between entering school and a year later – from the SEA to 6 year Net (Figure 4). Clearly, the bottom third of the SEA students barely changes over the first year, whereas those in the top two-thirds show remarkable gains by age 6.

Discussion and conclusions

An oft claimed recommendation is to provide a method so that gains, growth, value-added can be demonstrated between the children's performance when they enter school and their subsequent attainment. One of the keys to achieving this aim is to ensure that the SEA and any consequential tests measuring reading achievement are placed on the same underlying reading dimension. This is not difficult from a measurement perspective, and would seem essential given the aim of teachers is to help children make major achievement gains over the years of schooling.

It must be noted immediately that it is 'desirable' that the correlation between performance on the SEA and any later assessment be lower (but not too low) – as a major aim of the schooling system is to change achievement regardless of what children start with at Year 1, inculcate values of self-regulation

and effort, and turn children onto liking school subjects and having confidence they can achieve. If the correlation is too high it may indicate that schools add value in the same proportion to the level of skills a child has upon entry to school; and this seems both inequitable and difficult to justify.

There are many recommendations for revising the SEA. A major effort needs to be placed on improving the reporting from the items and scales to teachers. The validity of tests relates to both the quality of evidence relating to the accuracy of the scores but also to the appropriateness of the inferences and actions deriving from any assessment. So much time is currently spent in schools individually administering the SEA (and 6-year Net) but it is far from clear that teachers derive the consequential gains in modifying their teaching, informs them of the strengths and gaps, and assist in providing them with alternative resources and methods of teaching (which is often obtainable at the item level; Hattie, 2009). Other enhancements relate to the method of administering. As part of an internal exploration the asTTle team developed a version of CAP that could be administered by teachers on a PalmPilot (Lee, 2005). This led to an easier and more systematic administration of the test, immediate reporting for the teachers, more information at the sub-scale levels, the development of growth trajectories, and immediate comparison of the student to “like student” performance (plus automatic uploading to student data records, school files, and other purposes). This information could readily then be used to develop age expectations as well as “exposure to learning” expectations (as recommended by Clay, 2002, p. 47).

A major focus in the use of the information from such tests is that the information should have implications for immediate teaching and learning strategies to ensure ALL students are better able to undertake their school tasks – and particularly to ensure that the gaps evident when children start school are not maintained, or increased. The emphasis in any assessment should be on transforming teaching and learning; hence the importance of the “What next” tied directly to student’s current performance. That so many students who start school “behind” their peers languish in this state a year later should be of major concern for the sector. A major opportunity seems to be missed to change practices, use evidence to not only make a difference but to show when differences are being made.

More specifically, additional items may be needed in the CAP assessing word meaning, and the use of subscale scores relating to word ordering, concepts about words, the grammar of reading, and correct placement of text. It is worth considering other suggestions such as Goodman's (1981), who recommended scale scores for book orientation, concepts about whether print or pictures carry the text message, concepts about directionality of lines of print, page sequences and directionality of words, concepts about the relation between written and oral language, and concepts of words, letter’s capitals,

space, and punctuation. Furthermore, consideration needs to be given to creating scores related to Book handling, Directionality, Word boundary, Conventions, and Hierarchical concepts about print. These changes can be incorporated by reducing the number of items (from 5 to 3) about assessing orientation, and increasing the number of items (from 7 to 10) assessing concepts about words. There may be a case to consider Anderson et al.'s (2001, p. 44) claim to add another test, alongside CAP about word recognition, letter identification, and phonological awareness. There are many computer based tools of these skills that could readily be added to an e-asTTle version. Gilmore (2004) also reiterated these improvements and suggested adding the assessment of letter identification, phonological awareness, naming vocabulary skills, and word recognition skills to the SEA.

One of the attractions of Tell Me is that students choose the books to be retold. This is reasonable provided appropriate methodologies are implemented to ensure consistency of interpretation across the books. Clearly this study has shown that there are major differences, but the use of item response models could readily make the appropriate adjustments. Again this is best done by using administration methods such as Palm Pilots, mobile phone methods, and other computer based systems. It is noted that teachers in most NZ primary and high schools are already big users of IRT models (linear programming, IRT estimation, etc.) when they use the asTTle application (Hattie, Brown, and Keegan, 200x). Until these systems are available it seem cautious to restrict the choice to two books that have the least bias and most comparability (Great Times with Granddad, and Terri and the Rocket).

A criticism by some groups of CheckOut is that it does not measure number strategies, and of course it was not so intended. It does seem that the richness of reporting does not seem to justify the greater administration time. But processes are quite different from mathematics knowledge and understanding and could readily sit along side measures of student understandings of number.

The major criticism is that any use of teachers or students time to administer tests must be offset by the power and richness of quality reporting information that leads to confirmation or disconfirmation of current practices. The first years of schooling can make a profound difference to students chances in schooling and the current situation seems to be favouring those who already start ahead, and there is least changes in those most in need of accurate and immediate reporting of strengths and gaps. Overall, the quality of the current battery is very high as most changes are minor. The problems seem less related to the psychometric properties and more to the immediacy of reporting.

References

- Anderson, P., Lindsey, J., Schulz, W., Monseur, C., & Meiers, M. (2001). *An evaluation of the technical and methodological aspects of the School Entry Assessment Kit*. Commissioned Report to the Ministry of Education. Melbourne, Aus.: ACER.
- Clay, M.M. (2002). *An observation survey of early literacy achievement* (2nd edition). Auckland, New Zealand: Heinemann
- Dewar, S., & Telford, M. (2003). *A Study of the Use of School Entry Assessment (SEA) in Schools*. Research Report. Wellington, NZ: Ministry of Education, Research Division.
- Education Review Office (2000). Early Literacy and Numeracy: The Use of Assessment to Improve Programmes for Four to Six Year Olds. Report #3. Wellington, NZ: ERO.
- Entwisle, D.R., & Alexander, K.L. (1993), Entry into School: The beginning school transition and educational stratification in the United States. *Annual Review of Sociology*, 19, 401-423.
- Gilmore, A. (2004) *Towards a Conceptual Framework for School Entry Assessment*. Ministry of Education, Wellington NZ.
- Goodman, Y. (1981). Test review: Concepts About Print Test. *The Reading Teacher*, 34, 445-448
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications
- Hattie, J.A.C. (2009). *Visible learning: A synthesis of 800+ meta-analyses relating to student achievement*. Oxford, UK: Routledge.
- Hattie (2009, 2012) Hattie, J.A.C. (2012). *Visible learning for teachers. Maximizing impact on achievement*. Oxford, UK: Routledge.
- Hattie, J.A.C., Brown, G.T., & Keegan, P. (2005). A national teacher-managed, curriculum-based assessment system: Assessment Tools for Teaching & Learning (asTTle). *International Journal of Learning*, 10, 770-778.
- Lee, Y.Y. (2005). *Mobile devices and synchronization concepts: Research report for e-asTTle*. Technical Advisory Report, Visible Learning Labs, University of Auckland, New Zealand.
- Leeson, H. V., Brown, G. T. L., Hattie, J. A. C., Irving, S. E., Reddish, P., MacKay, A. J., & MacCall, S. (2005, April). *NZ students' performance in reading* (asTTle Monograph No. 3). Auckland, NZ: University of Auckland & Ministry of Education.
- McNaughton, S., (1995). *Patterns of Emergent Literacy: Processes of Development and Transition*. Auckland, NZ: Oxford University Press.
- Ministry of Education. (1997a). *School Entry Assessment Kit*. Wellington, NZ: Learning Media Ltd.

- Ministry of Education. (1997b). *SEA Guide to Teachers*. Wellington, NZ: Learning Media Ltd.
- Philips, D. (2000). Curriculum and assessment policy in New Zealand: Ten years of reforms. *Educational Review*, 52(2), 143-153.
- Prior, M., Bavin, E., & Ong, B. (2011). Predictors of school readiness in five- to six-year-old children from an Australian longitudinal community sample, *Educational Psychology*, 31(1), 3-16
- Rock, D.A., & Stenner, A.J. (2005). Assessment issues in the testing of children at school entry. *The Future of Children*, 15(1), 15-34.
- Young-Loveridge, J. (1997). From Research Tool to Classroom Assessment Device: The development of Checkout/Rapua, a shopping game to assess numeracy at school entry. In Biddulph, F. & Carr, K. (eds.) *People in Mathematics Education. Proceedings of the 20th Annual Conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp 608-614), MERGA, Rotorua, NZ.

Table 1. SEA Factor loadings on three factors

SEA Test & Scores	SEA Factors		
	1	2	3
<i>Checkout</i>			
Numerals	-0.02	0.86	-0.03
Patterns	-0.08	0.80	-0.04
Sets	0.04	0.82	0.03
Sequences	0.04	0.81	0.02
Operations	0.04	0.64	0.03
<i>Tell Me</i>			
Comprehension	0.46	0.24	0.09
Sentences	0.87	0.00	-0.01
Vocabulary	0.87	-0.01	0.01
Organisation	0.91	-0.03	-0.03
Description	0.90	-0.05	-0.03
Content	0.85	-0.01	0.01
<i>Concepts about Print</i>			

CAP: Language	-0.03	-0.14	0.92
CAP: Orientation	0.00	0.09	0.74
CAP: Word concepts	0.01	0.04	0.95
Correlations	1	2	3
<i>1. Checkout</i>	—		
<i>2. Tell Me</i>	0.40	—	
<i>3. Concepts About Print</i>	0.21	0.20	—

Table 2. Factor and Item Statistics for CAP

		Item Statistics			Factor pattern		
		Item- total					
No.	Item	<i>M</i>	<i>SD</i>	<i>r</i>	Word	Grammar	Placement
<u>Word ordering</u>							
I13	What's wrong: reordering of letters	0.03	0.16	0.14	0.64	-0.01	0.01
I12	What's wrong: word sequence	0.04	0.20	0.16	0.62	0.00	0.04
I14	Reordering of letters in a word	0.02	0.13	0.05	0.40	-0.04	-0.01
I20	Identify "was", "no"	0.07	0.26	0.19	0.29	0.13	0.02
I10	Line sequence incorrect	0.10	0.31	0.29	0.27	0.27	-0.01
<u>The grammar of reading</u>							
I23	Punctuation: quotation marks	0.33	0.47	0.56	-0.06	0.72	-0.0
I22	Where do I start reading?	0.31	0.46	0.55	-0.01	0.70	-0.06
I19	Show bottom of picture	0.34	0.47	0.52	0.00	0.58	0.01
I9	Capital letter concepts	0.41	0.49	0.53	0.05	0.52	0.08
I24	What is question mark?	0.24	0.43	0.42	0.01	0.51	-0.04
I21	Which way do I go?	0.60	0.49	0.52	-0.15	0.49	0.17
I7	Point to word as I read	0.53	0.50	0.58	0.01	0.48	0.20
I15	Punctuation: full stop	0.25	0.43	0.39	0.04	0.44	0.00
I16	Orientation of the book	0.52	0.50	0.50	-0.06	0.44	0.16
I6	First and last letter concepts	0.40	0.49	0.53	0.12	0.40	0.19
I8	Punctuation: comma	0.65	0.48	0.47	-0.06	0.33	0.25
I18	Letter concepts	0.10	0.30	0.23	0.04	0.25	-0.01
I17	Word concepts	0.07	0.25	0.20	0.11	0.21	-0.02
<u>Placement of text</u>							
I4	Response to inverted print	0.79	0.41	0.66	0.10	-0.11	0.99
I5	Capital and lower case letters	0.74	0.44	0.67	0.11	-0.01	0.88
I3	First and last part of story	0.81	0.39	0.64	0.04	-0.03	0.87
I2	Where do I begin?	0.85	0.36	0.56	-0.06	0.06	0.67
I1	Show me where to start	0.88	0.33	0.50	-0.11	0.14	0.51
I11	Where do I go after that?	0.69	0.46	0.47	-0.10	0.26	0.34

Table 3. Specific Language at Home Scores and Effect Size Differences to English

Language at Home	<i>N</i>	<i>M</i>	<i>SD</i>	Effect-size: English-
English	47334	10.04	4.89	
Fijian	33	8.79	3.52	0.30
Asian	261	8.72	5.46	0.25
Japanese	45	8.62	5.49	0.27
Korean	192	8.59	5.62	0.28
Chinese	834	8.68	5.15	0.27
Cantonese	219	8.76	4.76	0.26
Mandarin	226	8.52	5.35	0.30
Indian	717	8.66	4.86	0.28
SE Asian	246	8.01	4.75	0.42
English/Maori	245	7.87	4.15	0.48
Arabic	162	7.67	4.85	0.49
Maori	177	7.53	4.43	0.54
African	80	7.18	4.30	0.62
Pasifika				
Niuean	75	7.07	4.36	0.64
Samoan	1700	6.29	4.09	0.83
Tokelauan	53	6.06	3.40	0.96
Other PI	74	5.88	4.06	0.93
Cook Island	137	5.87	3.71	0.97
Tongan	751	5.79	3.74	0.99
Total	53140	9.75	4.93	

Table 4. CAP Scores and Effect Sizes by Ethnic Group

Ethnic Group	M	SD	No.	Effect-size
Pakeha	10.73	4.83	35807	Pakeha -
Maori	7.95	4.37	11800	-0.60
Pasifika	6.93	4.21	4812	-0.84
Samoan	7.17	4.38	2782	-0.77
Tongan	6.48	4.04	1060	-0.96
Other Pasifika	7.72	4.24	1650	-0.66
Asian	9.26	5.17	2577	-0.29
Chinese	9.35	5.33	1024	-0.27
Other Asian	9.18	5.02	1553	-0.32
Total	9.75	4.93	53140	

Table 5. IRT Item Values for CAP.

Number	Item	IRT Item Values	
		<i>b</i> (Difficulty)	<i>a</i> (Discrimination)
I1	Show me where to start	-1.41	1.53
I2	Where do I begin?	-1.18	2.03
I3	First and last part of story	-0.91	3.13
I4	Response to inverted print	-0.78	4.17
I11	Capital letter concepts	-0.77	0.82
I5	Where do I go after that?	-0.65	3.19
I8	Which way do I go?	-0.59	0.79
I21	Word concepts	-0.38	0.91
I7	Punctuation: comma	-0.09	1.18
I16	What is question mark?	-0.09	0.82
I9	Orientation of the book	0.34	1.01
I6	Capital and lower case letters	0.37	1.06
I23	What's wrong: reordering of letters	0.57	1.27
I19	Letter concepts	0.59	1.02
I22	What's wrong: word sequence	0.63	1.27
I24	Reordering of letters in a word	1.14	0.80
I15	Where do I start reading?	1.18	0.70
I10	Point to word as I read	2.17	0.73
I18	Line sequence incorrect	2.90	0.51
I20	Identify "was", "no"	3.20	0.53
I17	Punctuation: full stop	3.33	0.52
I12	First and last letter concepts	3.69	0.56
I13	Punctuation: quotation marks	3.93	0.62
I14	Show bottom of picture	10.88	0.22

Table 6. Ethnic Differences in Checkout Scores

Check Out Score	Ethnic Group							
	Cook					Other		
	Maori	Samoan	Island	Tonga	Other PI	Chinese	Asia	Other
Numerals	0.56	0.32	0.00	0.81	0.48	-0.30	-0.10	0.05
Patterns	0.42	0.29	0.24	0.56	0.39	-0.30	0.04	0.12
Sets	0.52	0.35	0.17	0.97	0.56	-0.10	0.10	0.18
Sequences	0.59	0.47	0.17	1.03	0.61	0.23	0.23	0.26
Operations	0.51	0.62	-0.10	0.82	0.59	0.02	0.23	0.25
effect-size: Pakeha-	0.52	0.41	0.09	0.84	0.53	-0.10	0.10	0.17

Table 7. Correlations between SEA and Six Year Net

SEA Scores	Six Year Net				
	Letters	Capitals	Words	Vocabulary	Dictation
CAP	0.88	0.99	0.78	0.75	0.91
Tell Me	0.21	0.24	0.32	0.23	0.24
Comprehension	0.55	0.57	0.48	0.51	0.59
Sentences	0.18	0.29	0.27	0.20	0.22
Vocabulary	0.03	0.16	0.13	0.01	0.05
Organisation	0.14	0.25	0.26	0.18	0.17
Description	0.07	0.19	0.21	0.14	0.12
Content	0.32	0.44	0.36	0.20	0.36
Check Out	0.35	0.32	0.49	0.48	0.37
Numerals	0.65	0.65	0.64	0.69	0.60
Patterns	-0.15	-0.11	-0.01	-0.02	-0.13
Sets	0.08	0.26	0.37	0.23	0.20
Sequences	-0.02	0.08	0.26	0.24	0.07
Operations	0.07	0.10	0.17	0.13	0.10

Table 8. Correlations between the SEA, PATs and asTTle Tests Administered About Four Years Later

SEA Scores	<u>Progressive Achievement Tests</u>			<u>asTTle</u>		
	Listening	Reading	Reading	Mathematic	Reading	Mathematic
	Comprehensio	Comprehensio	Vocabular			
	n	n	y	s		s
Concepts						
About Print	0.30	0.36	0.45	0.28	0.40	0.43
Tell Me	0.31	0.21	0.42	0.35	0.15	0.22
Comprehen						
sion	0.17	0.28	0.40	0.29	0.21	0.25
Sentences	0.38	0.67	0.79	0.33	0.39	0.16
Vocabulary	0.41	0.65	0.74	0.32	0.41	0.23
Organisatio						
n	0.34	0.67	0.79	0.27	0.35	0.21
Description	0.29	0.67	0.79	0.23	0.33	0.19
Content	0.12	0.36	0.65	-0.01	0.43	0.19
Check Out	0.41	0.31	0.52	0.52	0.26	0.44
Numerals	0.30	0.33	0.50	0.43	0.37	0.43
Patterns	0.04	0.07	-0.06	0.13	0.27	0.03
Sets	0.51	0.16	-0.02	0.62	0.08	0.34
Sequences	0.52	0.57	0.58	0.55	0.19	0.41
Operations	0.32	0.21	0.29	0.21	0.19	0.21

Table 9. SEA Scores by 6 Year Net Category One Year Later

6 Year Net	SEA Scores					
	Concepts about Print		Tell Me		Mathematics	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Letters						
1 (Low)	1.47	11.99	31.25	21.75	31.52	21.50
2 (Middle)	46.47	9.58	47.64	13.21	46.02	14.98
3 (High)	52.53	2.52	44.82	16.71	44.44	18.57
CAP						
1 (Low)	5.43	3.32	11.25	6.44	11.25	6.37
2 (Middle)	13.32	2.12	16.32	4.81	15.89	4.98
3 (High)	19.17	1.84	15.44	5.43	15.65	5.86
Word						
1 (Low)	4.36	2.59	7.01	4.31	6.95	4.25
2 (Middle)	7.34	3.86	10.06	4.24	9.23	3.93
3 (High)	11.77	3.02	10.20	3.95	11.17	4.14
Vocabulary						
1 (Low)	4.20	4.15	15.09	16.07	14.73	15.55
2 (Middle)	15.81	11.36	3.56	21.94	25.09	18.58
3 (High)	37.48	2.31	3.47	26.34	37.00	27.99
Dictation						
1 (Low)	6.63	5.91	17.46	12.59	17.34	12.36
2 (Middle)	21.13	8.31	27.16	1.70	25.42	11.09
3 (High)	32.51	4.84	25.98	1.93	27.87	11.43

Figure 1. Test Information Curve for showing information is high for assessing all students in the sample

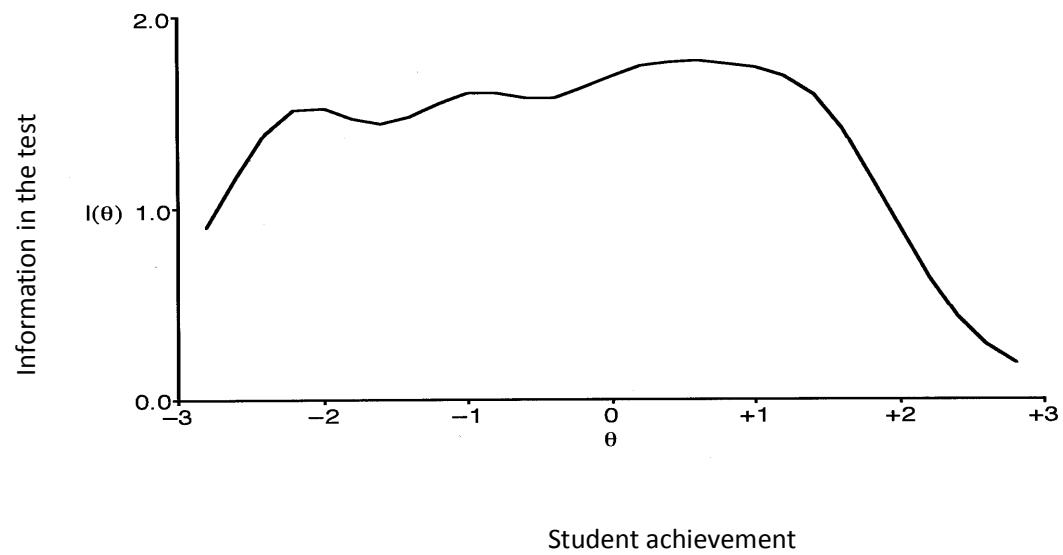


Figure 2. Mean Differences in Tell Me Scores and Stories

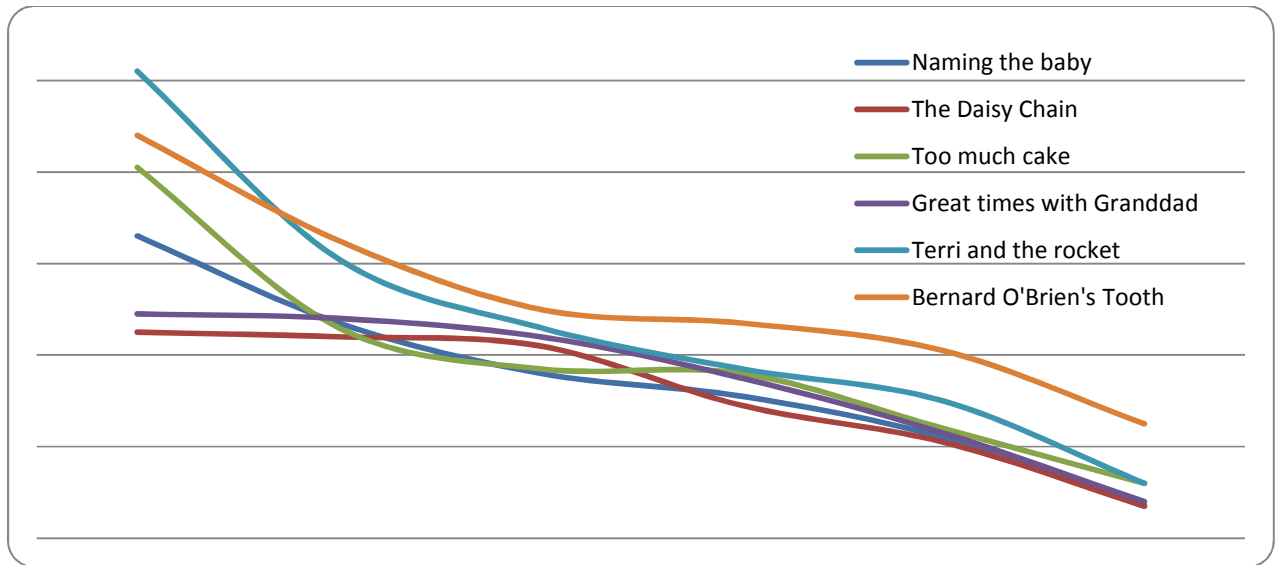


Figure 3. Means on the 6 dimensions of Tell Me by Ethnicity.

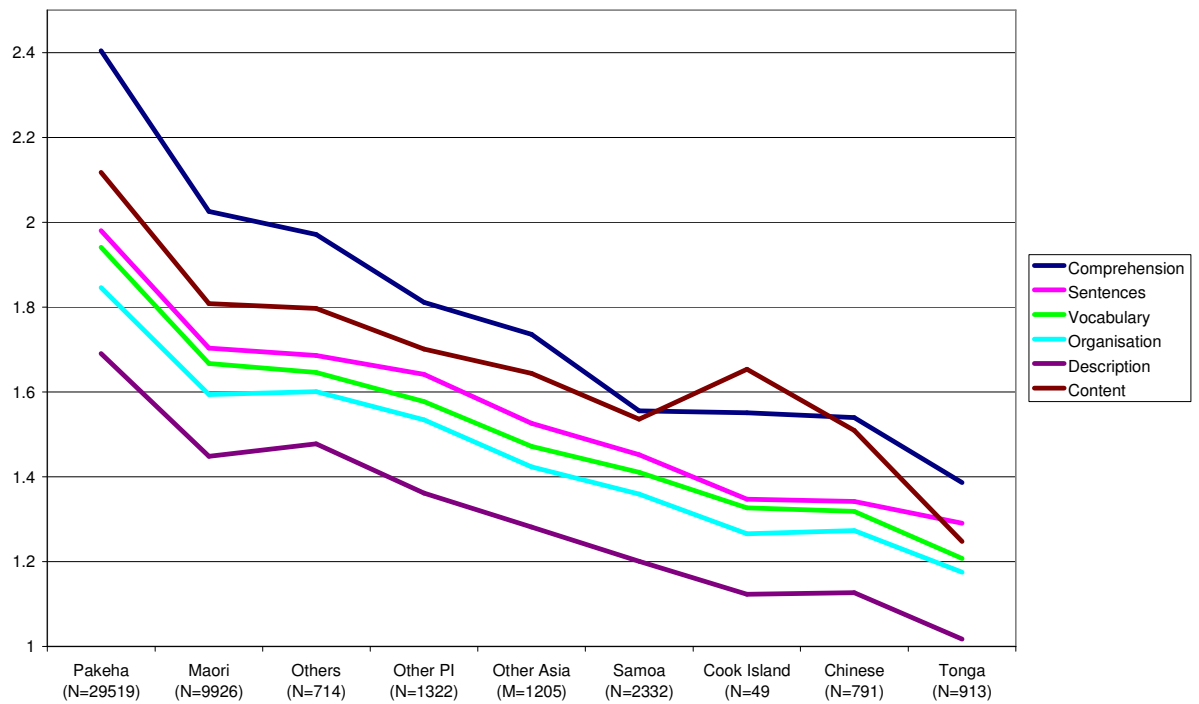


Figure 4. Relationship of Year 6 Net Scores to SEA Score Category

