Australian Indigenous students' performance on the PIPS-BLA Reading and Mathematics scales: 2011-2013

Irene Styles
The University of Western Australia

Helen Wildy
The University of Western Australia

Vivienne Pepper
The University of Western Australia

Joanne Faulkner
Murdoch University, Western Australia

Ye'Elah Berman
The University of Western Australia

Abstract

The assessment of literacy and numeracy skills of students as they enter school for the first time is not yet established nation-wide in Australia. However, a large proportion of primary schools have chosen to assess their starting students on the Performance Indicators in Primary Schools-Baseline Assessment (PIPS-BLA). This series of three studies aimed to establish whether (a) items in the Reading and Mathematics scales of the PIPS-BLA function differently for Indigenous and non-Indigenous students; (b) easier items developed for both scales are able to target Indigenous low-scoring students more reliably; and (c) factors such as gender, age, geolocation and ESL status are related to high and low levels of performance. The samples consisted of Indigenous and non-Indigenous students from metropolitan, regional and remote areas in Western Australia starting school in the years 2007 to 2013. The Rasch model was used to examine the operation of items. Both scales showed very little differential item functioning; the new items improved the measures for low-scoring students only marginally; and ESL status predicted performance most strongly. Female students performed better than male students and regional and metropolitan students better than remote students. These findings support previous research. Some reasons for the lowest performances are discussed.

Keywords

Assessment, early childhood, Indigenous, Rasch measurement

Introduction

In Australia, the academic performances of many Indigenous students, and their performance on standardised tests of achievement and ability have raised the concern of governments, institutions, and private individuals for many decades (Catholic Education Commission, 2000; Commonwealth Government, 1988, 1990; MCEETYA, 2000; WAACHS, 2006). This has been a concern common to several countries, including Canada, the United States and New Zealand, since it is seen to impact on individuals who are prevented from fulfilling their potential and achieving their most valued goals, and is, thereby, a great cost to the nation both economically and socially. This paper seeks to further understanding about the test performance and characteristics of Australia's youngest Indigenous students—those just beginning school—since educational and health

ISSN 1838-0689 online

research for many decades has pinpointed experiences in the early years of child development as vital to robust later development and adult life (for example, Australian Early Development Index National Report, 2012; Brooks-Gunn & Duncan, 1997; Moore, 2005; Wadsworth, 1999). To achieve our overall aim, we used data from the Performance Indicators in Primary Schools – Baseline Assessment (PIPS-BLA) in conjunction with Rasch measurement theory. An earlier paper (Godfrey & Galloway, 2004) evaluated the reliability of the PIPS-BLA for use with Indigenous students from the perspective of classical test theory, but noted that a larger sample would help confirm their findings. Earlier, Godfrey (2003) reported sound psychometric properties for the PIPS-BLA with a sample of Indigenous students. In the present study, we used the Rasch measurement model for dichotomous data to examine whether there is significant evidence of differential item functioning which may disadvantage Indigenous students, and, if not, what the performance for a larger sample of Indigenous students is, relative to that of non-Indigenous students. Rasch theory is considered by many researchers to be an advance on the classical perspective in terms of measurement principles (for example, Andrich, 1988; Rasch, 1960/1980; Wright, 1999).

Background

The performances of students at all levels of schooling in Australia have been assessed routinely on a range of literacy and numeracy knowledge and skills for more than a decade. Earlier assessments carried out by state educational bodies have recently been replaced by national assessment, the National Program on Literacy and Numeracy, for Years 3, 5, 7 and 9. The assessment of students as they enter school has not as yet been formalised through a national program of testing: instead, different states and different schools within each state have chosen different standardised assessment strategies, or none.

In Western Australia (as in most other states), many schools have opted to use the Performance Indicators in Primary Schools-Baseline Assessment (PIPS-BLA) which was developed at the Centre for Evaluation and Monitoring at Durham University in the United Kingdom. It identifies students' levels of cognitive development on entering school for the first time, and predicts later learning (Styles, 2009; Tymms, 1999; Tymms et al, 2007). It encompasses knowledge and skills that are considered essential for successful performance in schools in western or westernised cultures. It consists of three main scales – Reading (including test items on vocabulary, ideas about reading and beginning reading skills), Mathematics (including test items on identification of digits, counting, shapes, and beginning addition and subtraction skills), and Phonological Awareness. The PIPS-BLA is a performance-based computerized assessment in which the students themselves respond to questions based on a series of pictures using a form of adaptive testing in that a sequence of questions is stopped if a student scores three consecutive items incorrectly. The PIPS-BLA has been described as having "high test-retest reliability, high predictive ability, low cost and comprehensive individual, class and school level data analyses" (Wildy & Styles, 2008b, p1). An indication of the regard with which the PIPS is held is that the Department of Education in the Australian Capital Territory (ACT) selected the PIPS as the most appropriate assessment for the early years of schooling (Early Years Consultative Committee, 2005).

This paper focuses on the performances of Indigenous students on the PIPS-BLA and seeks to answer the question of whether such a test might be appropriate for this group of students, and what factors may be important in predicting performance.

Two major concerns in regard to the education of Indigenous students have been debated over many years. The first concern is the overall lower performance of Indigenous students in both primary and secondary schools across Australia compared with that of non-Indigenous students (Bradley et al, 2007; de Bortoli & Thompson, 2010; Harslett, 1996; Lokan, Ford & Greenwood, 1997; MCEECDYA, 2010; McInerney, 1991). Using the teacher-report Australian Early Childhood Index, the Australian Early Development Index National Report notes that Indigenous students in Australia have twice the rate of being developmentally at-risk than do non-Indigenous students. This is a concern Australia has in common with other westernized countries such as New Zealand, Canada and the United States. However, differences between Indigenous and non-Indigenous groups are not as marked in those countries as they are in Australia (WAACHS, 2006). As a result, many different commonwealth and state governmental, institutional and private strategies have been put in place over several decades in attempts to counteract this pervasive trend. Strategies have included additional health and educational resources to schools with high proportions of Indigenous students—and especially schools in remote areas of Australia where the difficulties and problems are exacerbated (see, for example, WAACHS, 2006).

Clearly, to evaluate and monitor performance, valid and reliable measures are required, and we expected that the PIPS-BLA may be a robust measure of performance in beginning literacy and numeracy, partly because of its established properties of validity and reliability, but also because of its format which involves one-on-one assessment with each student responding to computerised items presented visually in an engaging way. In contrast, other measures of performance in early childhood involve teachers' reports of their evaluations of students' performances. One such assessment tool is the Australian Early Development Index (AEDI) which has been shown to have robust psychometric properties for both Indigenous students (Silburn, Brinkman, Ferguson-Hill, Styles, Walker & Shepherd, 2009) and non-Indigenous students (Andrich and Styles, 2004). The concern in this paper is what Godfrey and Galloway (2004, p153) refer to as "the unfair, often subtle assessments made of Indigenous children by school personnel without the aid of reliable and valid instruments". Such biases may also influence teacher reports in some instances, even with the aid of standardized instruments, so tests which the students themselves perform are more likely to be accepted as valid. One of the questions we posed, therefore, was whether the results of our study would indicate a higher performance level for Indigenous students relative to non-Indigenous students than is typical on other measures. Another question which arose from the analyses of data in regard to the first question, was whether the spread of the item difficulties in two of the scales—Reading and Mathematics—could be increased at the lower end in order to measure low-scoring students more reliably.

The second major concern is about the suitability of tests based on skills and knowledge valued by westernized societies being used to assess students from other cultural backgrounds, and particularly those from Indigenous cultures (Klenowski, 2009; Luke, et al, 2002; Stobart, 2005; Tripony, 2002). Godfrey and Galloway (2004) report a wide range of views from Indigenous educationalists in Western Australia both in regard to the use of a particular test (Diagnostic Reading and Spelling Tests 1 and 2 (Waddington 2002)) which they had carefully researched for its appropriateness, and to the administration of standardized tests to Indigenous students generally. This led to their substituting the PIPS-BLA for their study, a choice that was regarded more favourably. Tests developed to assess performance in a westernised school system are often seen to be biased against students from other cultures in terms of the content and at least some of the skills addressed. The knowledge and skills represented in most such tests are those which have been deemed important in coping successfully in a western culture which, currently in Australia, is the dominant culture. And, arguably, any difficulties Indigenous students may face in obtaining and using such knowledge and skills are likely to disadvantage them in being successful in a broad range of contexts within the dominant culture (see, for example, Tripony, 2002), unpalatable as this fact may be. Thus another question we address is, given the context in which schools operate, whether the individual items which constitute the PIPS-BLA are biased against Indigenous students. In this paper, this question turns on the possible presence of differential item functioning (DIF), that is, whether Indigenous students have a lower chance of answering particular items correctly, even when these students have the same overall score as non-Indigenous students. We stress here that the question posed deals with knowledge and skills that are represented in a test such as PIPS, and not whether there are other skills or knowledge in which Indigenous students may out-perform non-Indigenous students but which have, at least to date, not been recognized as essential in the dominant culture.

A final question was what profiles of demographic factors are associated with, firstly, Indigenous students with the highest performances and, secondly, those students with the lowest performances. The demographic factors available to be investigated were those routinely recorded by teachers as part of the computerized administration of the PIPS-BLA, namely, sex (male or female); age at assessment; geographical location (metropolitan, regional or remote areas in Australia); and English as a second language (ESL) status (Yes or No). In studies involving the AEDI, significantly lower performances have been noted for ESL, remote, and male students (AEDI National Report, 2012).

In the research reported in this paper, three studies with different aims were carried out. The first study examined the important question of whether the PIPS-BLA Reading and Mathematics scales are appropriate for assessing Indigenous students in terms of whether individual items within the context of a whole scale may be disadvantaging Indigenous students: this is a question of whether some items exhibited differential item functioning, that is, operated differently for Indigenous and non-Indigenous students. Only once the suitability of items was established could we then legitimately compare the performances of Indigenous and non-Indigenous students. The second study checked whether new items which were developed to assess the lowest-scoring students more reliably were successful in doing this. The third study sought to identify some of the demographic characteristics of the lowest- and highest-scoring Indigenous students. Methods used to address the questions for each study are described in the following three sections.

Study 1: the suitability of scales for assessing Indigenous students

This study addressed two questions. The first was whether items in the two scales (Reading and Mathematics) showed evidence of differential item functioning, that is, were operating differently for Indigenous and non-Indigenous students. Another perspective on this question is whether the scales represent the same constructs for both groups of students: if they do not, then the performance of the two groups should not be compared. The second question was, if the appropriateness of the scales was established, what were the overall levels of performance of Indigenous students compared with those of non-Indigenous students, and, in particular, were the Indigenous performance levels relatively higher than those apparent in studies using other measures.

Responses to the PIPS-BLA Reading and Mathematics scales from students starting school for the first time in the years 2007, 2008 and 2009 in Western Australia were collected from Independent, Catholic and Government schools—a total of 32 315 students. Then all Indigenous students (n=1

373) in this sample, plus a random sample of approximately the same number of non-Indigenous students (n=1 381) from the remaining subset of students, were selected for this analysis.

To address the first question, Rasch measurement theory (RMT) data (see for example, Rasch, 1960/80; Andrich, 1988) was used to examine the properties of the scales and to provide person measures (or locations) which could be subjected to standard ANOVA techniques to address the second question. A major advantage of using Rasch theory is the provision of a linear scale which is invariant within the frame of reference for which it has been checked. Rasch (1961) rendered these characteristics of measurement in terms of invariance of comparisons within a specified frame of reference defined by a class of persons responding to a class of items, as follows:

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; …
>
> Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison; … (Rasch, 1961, p. 322).

In social science measurement, the stimuli are generally referred to as items, where a set of items form an assessment or a test.

The class of models that is generated by the above requirements of invariance is characterized by sufficient statistics for the person and item parameters. Sufficient statistics permit the estimation of the item parameters independently of the values and distribution of the person parameters.

Rasch measurement theory essentially involves transforming raw scores on a test into linearized measures, or locations, for both persons and items. The Rasch model (RM) for dichotomous responses, which arises from the above requirements of invariance, takes the form

$$\Pr\{X_{ni} = x\} = [\exp(x(\beta_n - \delta_i))] / \gamma_{ni}, \qquad (1)$$

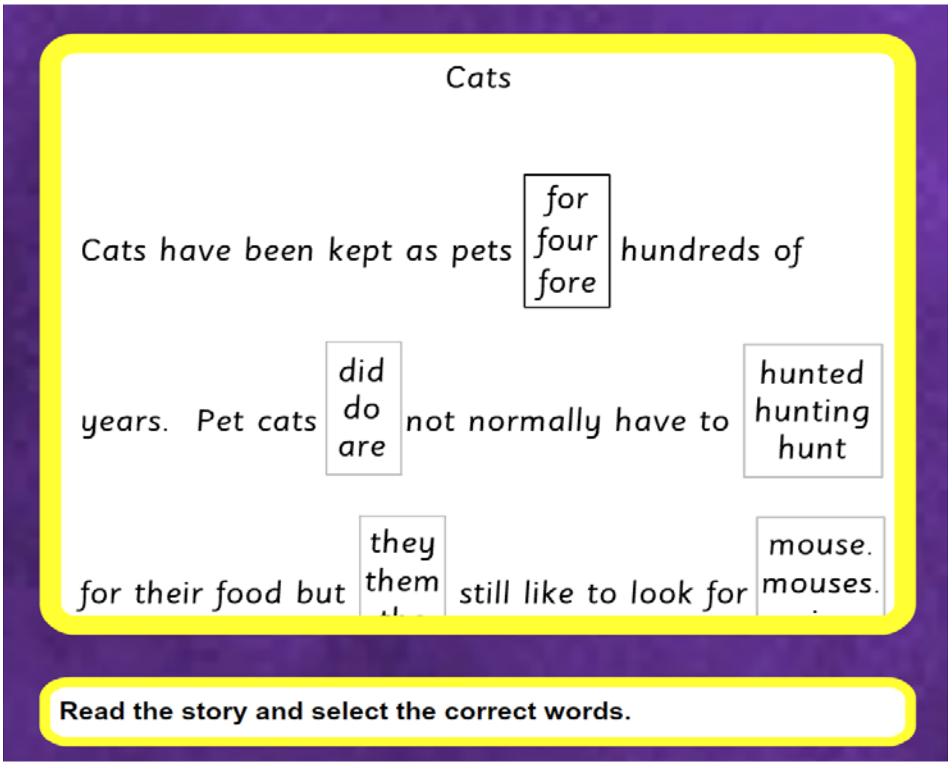
where the random variable $X_{ni} = x$, $x \in \{0.1\}$ characterizes the incorrect and correct responses, β_n and δ_i are respectively the proficiency and difficulty of person n and item i, and $\gamma_{ni} = 1 + \exp(\beta_n - \delta_i)$ is the normalizing factor ensuring that the sum of the probabilities sums to 1.

The software RUMM2030 (Andrich, Sheridan & Luo, 2008), incorporating a wide range of facilities with which to examine the operation of sets of items, was used for the Rasch analyses.

Results

Reading scale

This scale consisted of 129 items covering aspects such as vocabulary, ideas about reading, and beginning reading skills. Four items had response data which were too extreme to be included in analyses (Cats 3, 4, 9, and 15). In this case, these items, which are among the most difficult in the entire set, were not answered correctly by any student, and since no comparative information was available for them, they could not be included in the Rasch analysis. Examples of these items are shown in Figure 1, where a student is asked to choose which of three options is the correct word to use in a sentence.



Cats

Cats have been kept as pets | for
four
fore | hundreds of

years. Pet cats | did
do
are | not normally have to | hunted
hunting
hunt

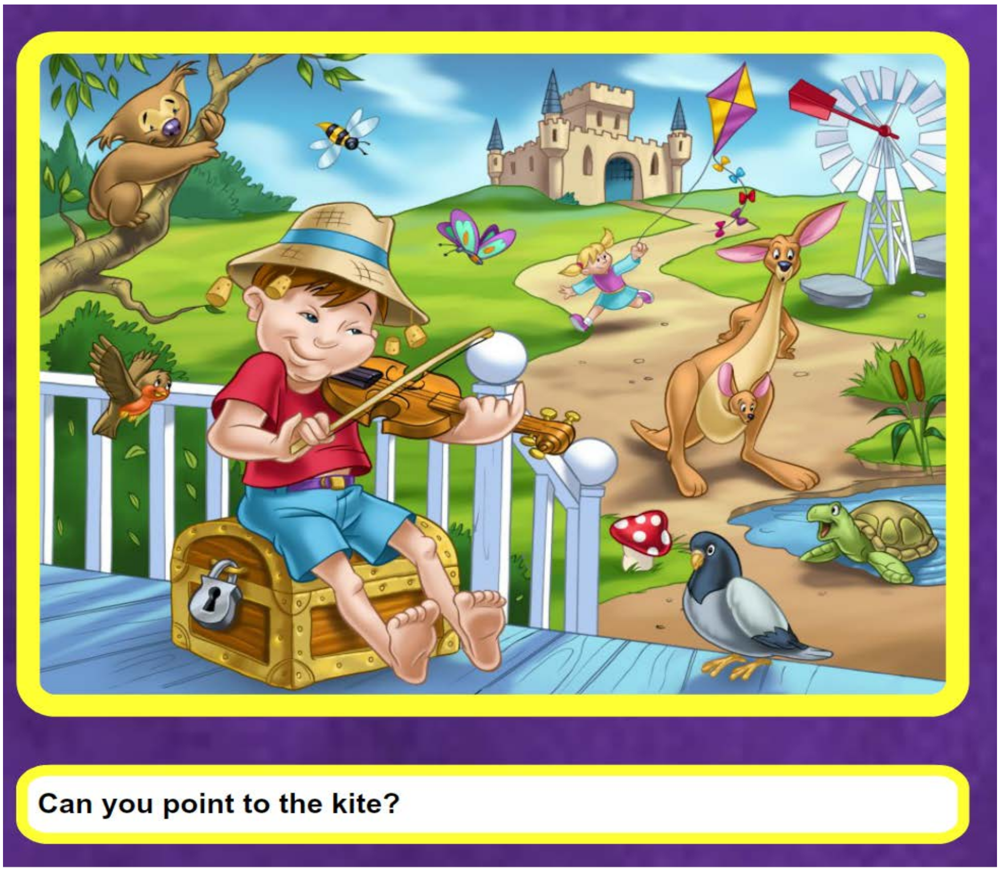
for their food but | they
them | still like to look for | mouse.
mouses.

Read the story and select the correct words.

Figure 1. Example of a Cats item from the PIPS-BLA Reading scale.

The good fit of the items to the Rasch model has been established in other studies (Wildy & Styles, 2008a). This was supported in the present study: when two items (identifying a *padlock* and recognizing the letter *s*) that fitted the Rasch model least well were removed, the fit of the remaining items, as judged by the statistical item-trait interaction and log residual tests of fit, and by the graphical test of fit using item characteristics curves, was acceptable. (Refer to, for example, Andrich (1988) for detailed explanations of these criteria).

Five items (recognizing the letter *s*, identifying the *knife* in a picture, *writing their own name*, identifying a *violin* and a *kite* in a picture) showed Differential Item Functioning (DIF) by Indigenous status. Two of these items, *kite* and *knife*, are shown in Figure 2.



Can you point to the kite?

Figure 2. Example of a Vocabulary item from the PIPS-BLA Reading scale.

DIF means these items operated differently for the two groups (Indigenous and non-Indigenous), with one group likely to perform better than the other group *even though students have the same total score* (Harquist & Andrich, 2004). Non-Indigenous students were more likely to get these items correct, though the reasons why this should be so are not obvious for the first three items. The difference on the last two vocabulary items is easier to interpret: it may be that Indigenous students (especially if they live in Remote communities) are less likely to have come across these objects. It is important to note that there was no DIF according to Indigenous Status for the highest scoring groups on any of these items. This means that items operate in a similar way for non-Indigenous and for higher-performing Indigenous students, but a few items operate differently for non-Indigenous and for lower-scoring Indigenous students, with the former scoring a little higher than the latter. These items were deleted from the analyses one by one on the basis of degree of DIF, until no further significant DIF was evident. The DIF items and the direction of DIF are shown in Table 1. After these items were removed, there was no DIF according to Indigenous Status or Gender. An alternative course of action could be to split the items into two separate items, one for each group (Harquist & Andrich, 2004). However, as there was no necessity to retain as many items as possible in these analyses, these items were simply omitted from further analyses.

Table 1. Items showing DIF according to Indigenous Status.

DIF factor	Item	Content	Direction
Indigenous Status	Letters	*s*	non-Indig > Indig
	Vocabulary	*knife*	non-Indig>Indig
	Name	*Write own name*	non-Indig>Indig
	Vocabulary	*violin*	non-Indig> Indig
	Vocabulary	*kite*	non-Indig> Indig

It is a significant result that all except a very few items may be accepted as measuring the same variable for both Indigenous and non-Indigenous groups—a variable whose items represent knowledge and skills which are deemed important in performing successfully in a westernised school environment. This means that the performances of both groups on this scale (without the DIF items) may legitimately be compared, bearing this frame of reference in mind.

The relative positions of item difficulties and person proficiencies which are located on the same continuum are referred to in Rasch theory as *locations* and the units of measurement as *logits*. (A logit is not a standard unit but may have a different origin and spread in analyses of different sets of items, and in every analysis the mean of all item difficulties is set at zero logits.) The distributions of item and person locations by Indigenous Status are shown in Figure 3 (n=1 386 for each group). It may be seen that the items are well-targeted to the students (most items are neither too difficult nor too easy for the majority of students), and that the scale allows for measurement of development over time (students are reassessed at the end of their first year at school) for virtually all students in this sample. A few students with very low locations are not being measured as reliably as most other students: this is because there are either some gaps in the item locations (for example, around -5 logits), or students are located below the easiest item location which is at -6 logits. Because of this pattern, a decision was taken to develop easier items which would target lower-scoring students so that they may assessed more reliably, thus increasing the discrimination amongst them. The results of doing this are presented in Study 2.

Reliability was very good with a Person Separation Index of 0.89 (the Rasch equivalent of Cronbach's alpha statistic). On average, Indigenous students did not perform as well as non-Indigenous students, with a mean difference equivalent to about one standard deviation. This finding is similar to those in the 2012 National Assessment Program in Literacy and Numeracy (NAPLAN) report on Year 3 students (http:www.nap.edu.au/verve_resources/naplan_2012_national_report.pdf). Note that Indigenous students are more variable in performance than non-Indigenous students, that is, they are spread further at both the lower and upper ends of the scale. There are thus likely to be proportionately more Indigenous students located at both the high and low extremes of the distribution of person locations. Many reasons have been put forward to account for these differences, including different cultural values, knowledge and skills (see for example, Luke et al, 2002; Stobart, 2005) and health problems (Leach, 1999) and many programs have been instituted to try to overcome educational difficulties in Indigenous communities (for example, MCEEDYA).

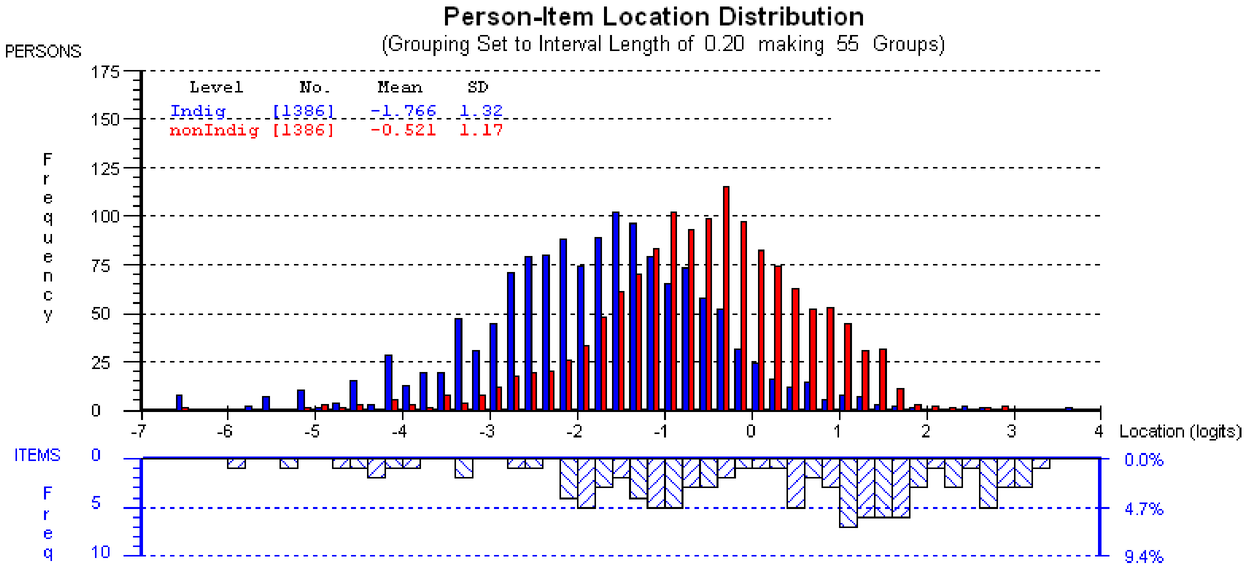


Figure 3. Distribution of items and person on Reading scale, according to Indigenous Status.

Mathematics scale

There were 69 items in this scale which includes items such as recognition of numbers, ideas about mathematics, and use of arithmetical operations. Four items (what is *42-17*, what is *a quarter of 8*), what is *twice 3 doubled*, and what is *105+302*) had extreme scores and had to be deleted from analyses. This was because no students attempted these items which are among the most difficult in the set and, since no comparative data were available for them, they were deleted from the Rasch analysis. As in the Reading scale, and similar to the findings in Wildy and Styles (2008a), the items fitted the Rasch model, once three items (*can you point to a hexagon*, *which is the shortest person?* and *how many fish are there?*) which did not show good fit to the model, had been removed.
Four items showed DIF by Indigenous status – *which bottle holds the least water?*, and recognizing the numbers *3*, *4* and *1*. The first of these items is shown in Figure 4.

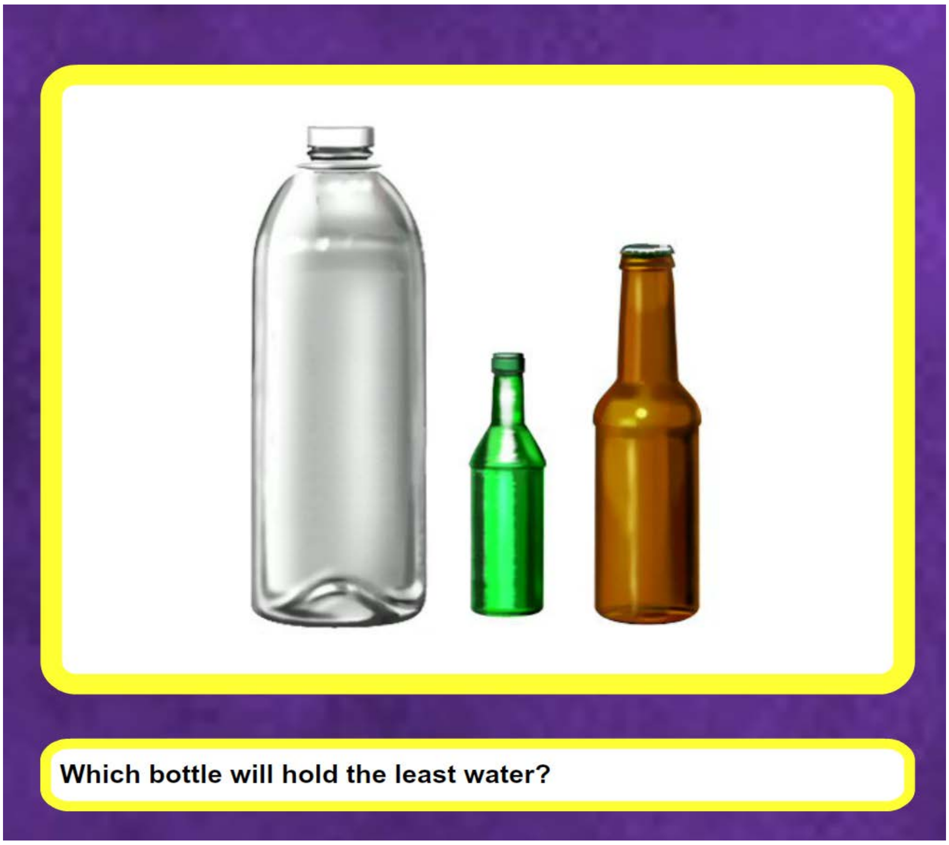


Figure 4. Example of an Ideas about Mathematics item from the PIPS-BLA Mathematics scale.

These items tended to operate differently for the Indigenous and non-Indigenous groups across the range of abilities except for the most able group of students where there was no DIF. The first item showed Indigenous students performing better than non-Indigenous students while the other three items showed the reverse. Table 2 lists these items and the direction of DIF. When the item *which bottle holds the least water?* was deleted from the analysis (it also showed poor fit to the model), no further significant DIF was evident and there was also no DIF according to Gender. The items requiring recognition of the numbers 4, 1 and 3 showed very small DIF across only some of the six total score groups – in practice the differences would not impact on total scores and so these items were retained in analyses.

Table 2. Mathematics items showing DIF according to Indigenous Status
.

DIF factor	Item	Content	Direction
Indigenous Status	Ideas about maths	*which bottle holds the least water?*	Indig>non-Indig
	Numbers	*3*	Non-Indig>Indig
	Numbers	*4*	Non-Indig>Indig
	Numbers	*1*	Non-Indig Indig>

From this analysis, we conclude that, as was the case for the Reading scale, the Mathematics scale, with the exception of four items, operates in a similar fashion for both Indigenous and non-Indigenous groups, that is, it measures the same variable for both groups. This means students from both groups are being measured on a construct which has the same qualitative meaning for both groups, and thus the performance of the two groups may legitimately be compared.

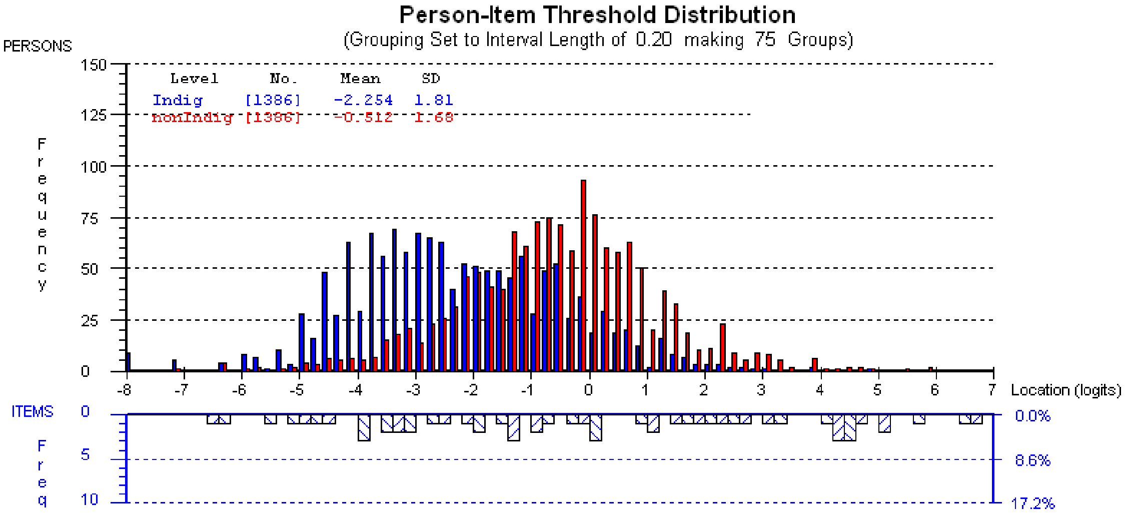


Figure 5. Distribution of items and person locations on the Mathematics scale according to Indigenous Status.

Figure 5 shows the distribution of item and person locations on the Mathematics scale by Indigenous Status. The Mathematics items are well-targetted to this sample of students (n=1 386 for each group) and the scale allows for improvement/development to be measured well for virtually all students. A small number of students at the lowest locations is not being measured as reliably as the other students. As with the Reading scale, there are either gaps in the item locations, or students are located below the easiest item location. For this reason, a decision was taken to develop five easier items to target these lower-scoring students more reliably. As with the Reading scale, Indigenous students on average did not perform as well as non-Indigenous students, with a mean difference equivalent, again, to about one standard deviation. Further findings on performance levels are presented in Study 3. This difference supports findings from other studies in the literature already mentioned.

Study 2: the performance of new items

In order to target low-scoring students, items theoretically easier than those in the original PIPS scales were developed – five for each of the Reading and Mathematics scales—and included in the PIPS-BLA. The analyses in this study used data from 656 Indigenous students starting school in Western Australia for the first time in 2011, 2012 and 2013.

Reading scale

For this analysis, the data for the sub-scales Stories (pictures with short sentences to read), Walking to School and Cats (both consisting of items similar to those shown in Figure 1) were not included because the last two scales had no data for this sample and the Stories had a large amount of missing data.

The overall fit of the items was satisfactory as evidenced by the mean log residual test of fit of 0.062 (sd=1.334) (a mean of close to 0 indicates good overall fit). The fit of the five new items to the Rasch model was relatively good in the context of the complex variable of Reading. *Can you point to the chair*? fitted the least well of the new items due to having relatively low discrimination. Surprisingly, however, the new items are not as easy as expected and none are easier than the original items. The easiest of the new items (*Can you point to the ball*) is second in order of increasing difficulty of all items and the next easiest new item (*Can you point to the cloud*) is fifth easiest overall. One item showed significant DIF by Gender (recognition of the letter *c*) where female students tended to perform better than males even though they had the same total scores) and one item exhibited DIF by Geolocation (knowing where the start of a sentence is) where Remote students tended to perform better than Regional—except for persons in the middle locations—and Regional students better than Metropolitan students across most locations. It is not clear why this was the case.

The Person Separation Index for all items was 0.916. The distributions of item and person locations with new items included and excluded are presented in Figure 6 and Figure 7, respectively. It is noted again that the origins (0 logits) of the item difficulties are different for the two sets of items. With the new items, the mean and standard deviation of the item locations were 0 and 2.098, respectively, with four students (0.6%) located below the easiest item and 13 students (2%) having just one item or no items below them. With the new items excluded, the item locations' mean and standard deviation were 0 and 1.964, respectively (indicating a greater spread of locations), there were no students below the easiest item and 18 (2.7%) students with just one item below their location. The mean standard error for the 12 lowest-scoring students reduced from 1.217 without the new items to 1.007 with the new items. Therefore, the new items have slightly improved the targeting of low-scoring students (five fewer students have just one item below their location), and the standard error of measurement has improved. It may be that there is consistently across the years a small group of students who perhaps do not understand the instructions, or are not motivated to respond well, rather than items being too difficult for them. Anecdotal evidence available from teachers for 11 students suggest this was the case for them. It may also be that some students have not yet learned the correspondence between the spoken word and pictorial representations of words. To investigate this further, in future administrations teachers will be asked to record their comments on the test-taking and in-class learning behaviours of low-scoring students.

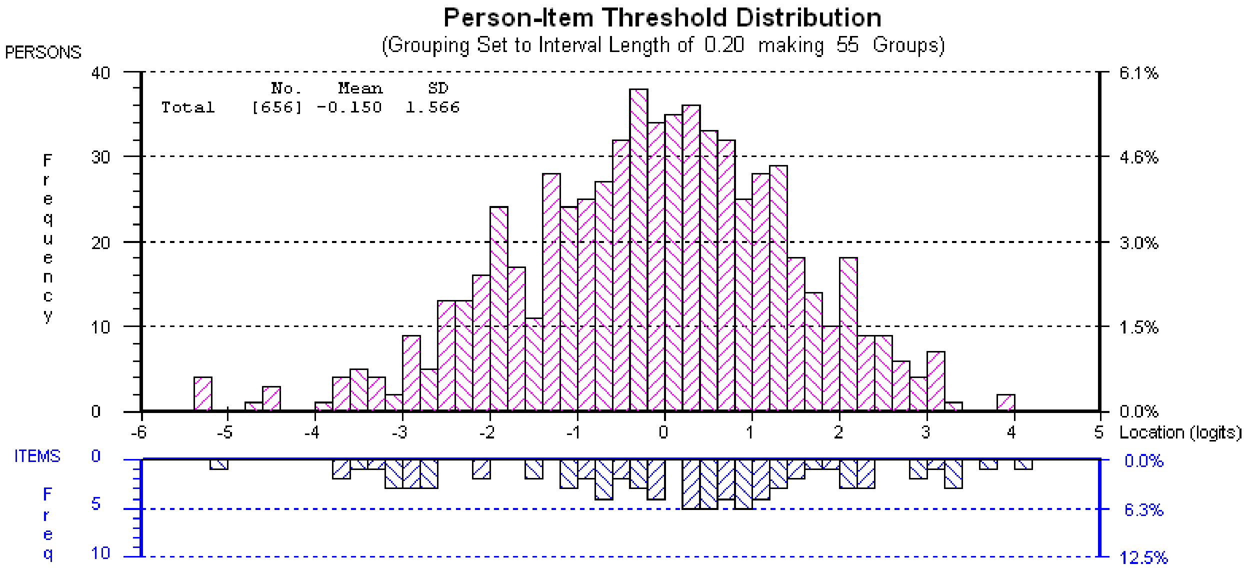


Figure 6. Distribution of person and item locations on Reading: new items included.

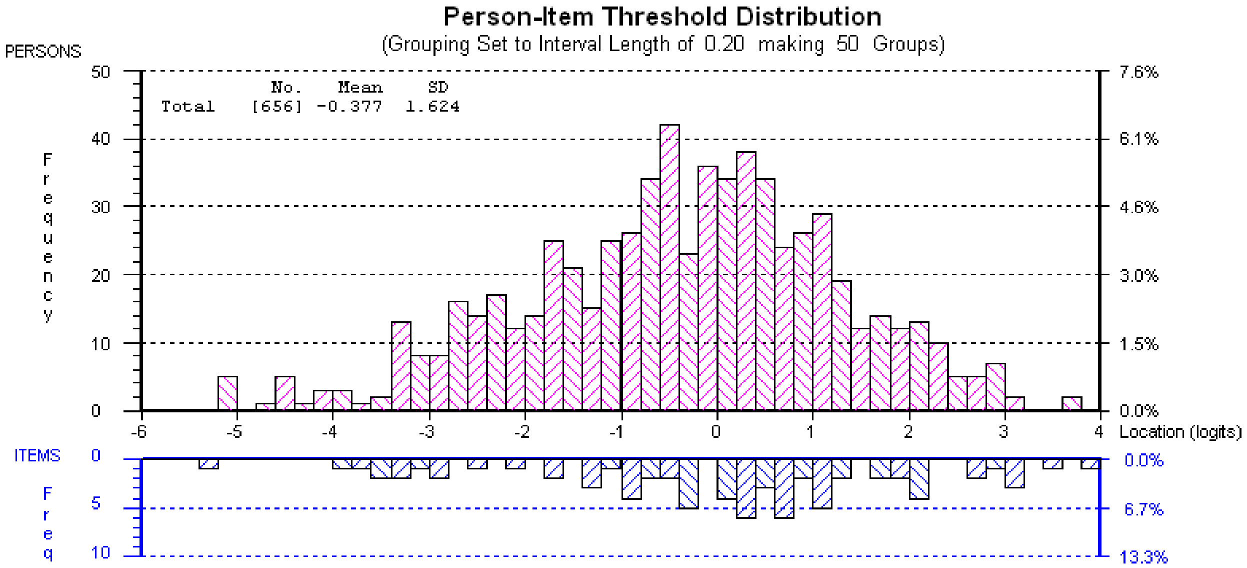


Figure 7. Distribution of person and item locations on Reading: new items excluded.

Mathematics scale

Overall fit of the items was good, as indicated by a mean log residual test of fit statistic of -0.292 (sd=1.361). For this scale the five new items fitted the Rasch model reasonably well but, again, they are not as easy as expected and none is easier than the original items. The easiest of the new items (*point to* a *biscuit that is on a plate*) is in third position of increasing difficulty. There was no significant DIF according to Gender and just one item showing DIF according to Geolocation (*31 balls*) where students from Regional backgrounds tended to score higher than those from Metropolitan or Remote areas even though they had the same total scores). There is no obvious reason why this is so. The Person Separation Index (reliability) for all items was a high 0.934.

To examine the impact of the new items on measuring the lowest-scoring students, the item-person distribution graphs for this sample, with and without the five new items included, are presented in Figure 8 and Figure 9.

ISSN 1838-0689 online

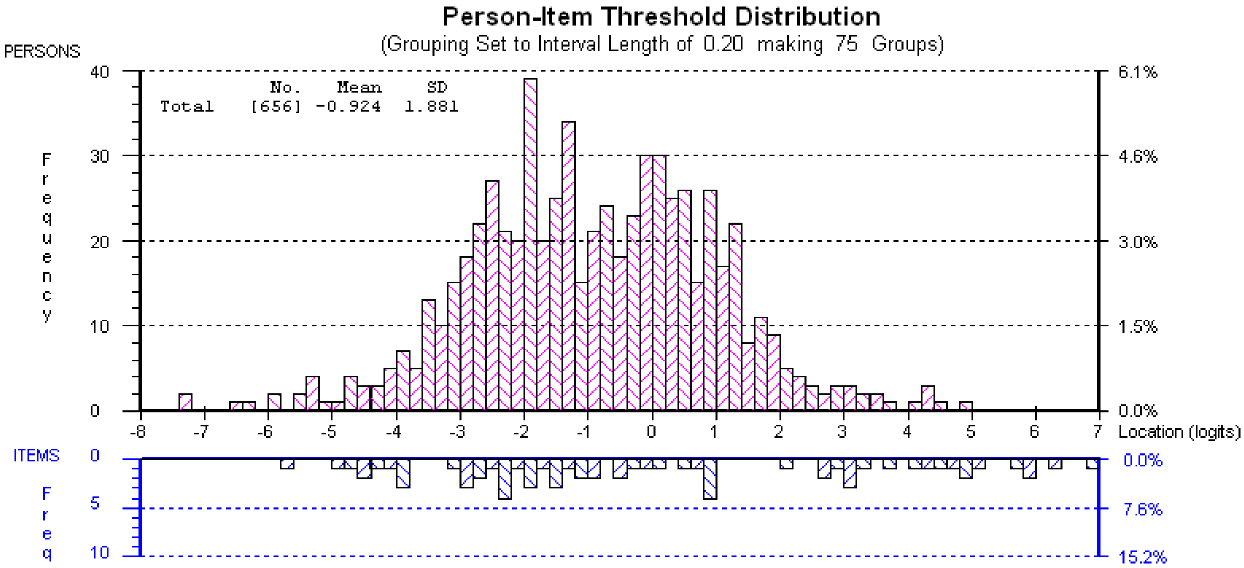


Figure 8. Distributions of item and person locations for Mathematics: New items included.

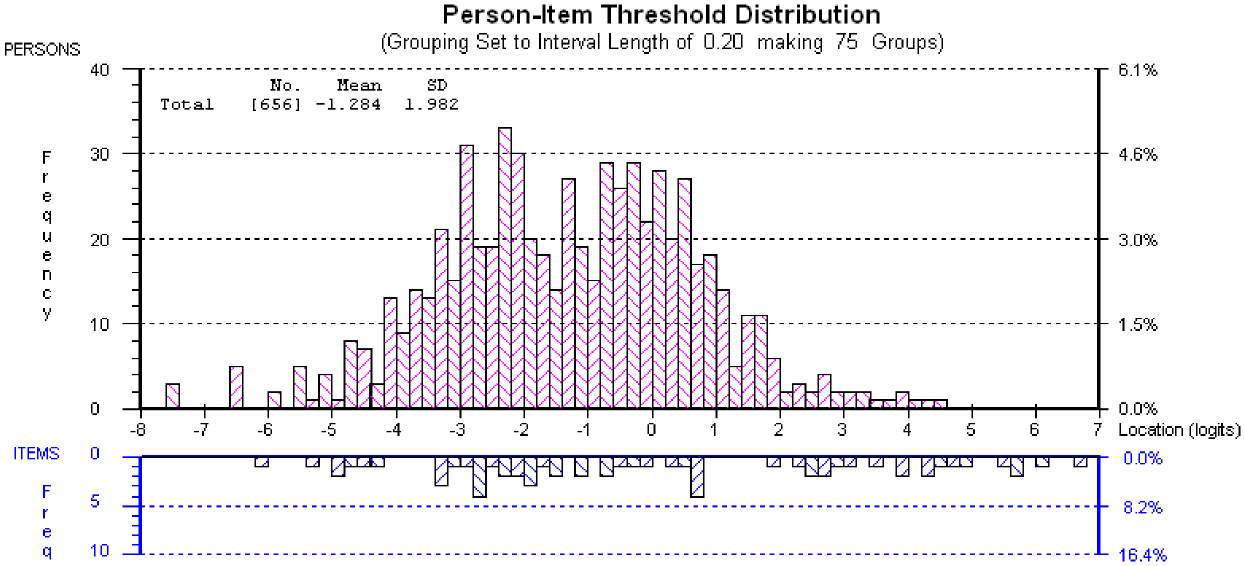


Figure 9. Distributions of item and person locations for Mathematics: New items excluded.

It can be seen that without the new items, the mean and standard deviation of the item locations was 0 (sd=3.374), with eight students (1.2%) below the easiest item which is at about -6.1 logits and 14 (2.1%) below the next easiest item. With the new items included, the mean of the item locations was 0 (sd=3.422) (showing a slight increase in item location spread), there are six students (0.9%) below the easiest item and 16 (2.5%) below the next easiest item. The mean standard error of measurement for the 12 lowest-scoring students reduced from 1.108 without the new items to 0.853 with the new items. The conclusion is that with the five new items, the Mathematics scale for PIPS-BLA has targetted the lowest-scoring students about the same, though there are two less students below the easiest item, and the standard error of measurement has improved. It is surprising that there are still a few students scoring below the easiest items: as suggested for the Reading scale, it may be that these students do not understand the test instructions or are not motivated to respond. And, again, there may be difficulties with recognizing the correspondence between spoken words and pictorial representations of those words. This is an issue that will be investigated further by asking teachers to provide comments on low-scoring students' test behaviours.

Aside from these concerns, there are some gaps in the item locations at other points on the continuum in Figure 10—for example between 1 and 2 logits, -3 and -4 logits and between -5 and -6 logits. It may be possible to develop items to fill these gaps and thereby provide more reliable measures for students who are located at these positions.

Study 3: Performance of Indigenous students

Reading scale

In order to compare the mean performances of Indigenous students with different demographic characteristics, the Rasch person locations (in logits) were analysed using standard one-way analyses of variance (ANOVA) techniques. On the Reading scale, firstly, the performances of students were, on average, different according to where they are located – students from Remote areas performed significantly less well (mean=-0.722 logits, sd=1.610) than students from either Regional (mean=0.244 logits, sd=1.390) or Metropolitan (mean=0.304 logits, sd=1.660) areas (F(2,653)=32.650, p=0.000005). (Note that the probabilities reported here are those available in the software: they are retained for the reader to judge the significance of the differences). There was also a significant difference (F(1.654)=34.928, p<0.000000) in means between the two Gender groups, with females out-performing males (mean=0.150 logits, sd=1.570 and mean=-0.603 logits, sd=1.690, respectively). Age-related differences were also significant (F(5,649)=4.880, p=0.0002), with performance increasing with increasing age from less than 4 years to 5.5 years, after which the means tended to decrease again. There was no significant difference in performance amongst Year groups (F(6,649)=0.505, p=0.805) though the Year 2011 cohort performed less well (mean=-0.382 logits, sd=1.760) than either the 2012 (mean=-0.133 logits, sd=1.650) or 2013 (mean=-0.145 logits, sd=1.60) cohorts. There was a significant difference between ESL and non-ESL groups (F(1,643)=143.334, p<0.0000005) with ESL students performing less well (mean=-1.387 logits, sd=1.540) than non-ESL students (mean=0.217 logits (sd=1.50)).

Mathematics scale

The ANOVA results for the Mathematics scale according to Gender, Geolocation, Age and ESL groups showed statistically significant differences in mean locations according to Gender, Geolocation, and ESL status. Females performed significantly better than males (mean=-0.776 logits, sd=1.810 and mean=-1.218 logits, sd=2.060, respectively) (F(1.653)=8.565, p=0.0035), Non-ESL better than ESL students (mean=-0.489 logits, sd=1.780 and mean=-2.362 logits, sd=1.670, respectively) (F(1,643)=144.637, p<0.000000), and Regional (mean=-0.436 logits, sd=1.720) and Metropolitan students (mean=-0.388 logits, sd=1.880) better than Remote students (mean=-1.614 logits, sd=1.860) (F(2.653)=37.357, p<0.0000005). The pattern of Age group means increases and then decreases across six age groups from 4 to 7 years of age. The decrease in the means which occurred for the three oldest age groups may reflect the fact that students starting school at these ages may have developmental difficulties which led to their later start.

The correlation between scores on the Mathematics and Reading scales was a high 0.784 (significant at p<0.01).

High-scoring students

Analysis of the top 1% and 5% of this sample of students (n=656) showed that the typical high-performing Indigenous student in Reading is likely to be female, aged between 5 and 6 years, with a non-ESL background, and living in a Metropolitan area. A typical high-performing Indigenous student in Mathematics is likely to be male, and also aged between 5 and 6 years, with a non-ESL background, and living in a Metropolitan area.

Low-scoring students

The most typical student in the lowest-performing groups (either the lowest 1% or lowest 5% of all students) on the Reading scale was male, aged five to five and a half years, of either ESL or non-ESL background, and living in either a Metropolitan or a Remote area. On the Mathematics scale, a typical low-scoring student was male, less than five years old, with either an ESL or a non-ESL background, and living in a Remote area.

Regression analyses

A regression analysis with Reading as the dependent variable and Gender, Geolocation, Age group, and ESL status as independent variables is shown in Table 3. Clearly, as evidenced by the beta weightings for each demographic characteristic, ESL status predicts performance most strongly, followed by Gender, then Geolocation, and, lastly, Age.

The beta weightings in Table 4 shows that with performance on Mathematics is predicted, again, mostly strongly by ESL status, followed by Geolocation, then Age, and Gender. Clearly, ESL status is an important factor in the performance on both the Reading and Mathematics scales of the PIPS-BLA.

Table 3. Regression analysis with Reading performance as the dependent variable.

Model	Unstandardised Coefficients		Standardized Coefficients		
	B	Std error	Beta	t	Probability level
(Constant)	-2.352	0.396		-5.936	0.000
Geolocation	-0.345	0.070	-0.186	-4.923	0.000
Age Group	0.205	0.084	0.086	2.433	0.015
Gender	0.709	0.118	0.212	6.023	0.000
ESL status	0.756	0.112	0.255	6.756	0.000

Table 4. Regression analysis with Mathematics performance as the dependent variable.

Model	Unstandardised Coefficients		Standardized Coefficients		
	B	Std error	Beta	t	Probability level
(Constant)	-2.615	0.447		-5.847	.000
Geolocation	-0.450	0.083	-0.208	-5.418	.000
Age Group	0.201	0.061	0.118	3.275	.001
Gender	0.406	0.137	0.106	2.965	.003
ESL	0.801	0.132	0.232	6.047	.000

Discussion

The findings from the analyses of these data are in line with those of other studies of Indigenous students' performance in schools.

In regard to the first question about possible bias in individual items, we demonstrated that most items in the PIPS-BLA Reading and Mathematics scales represent, and thus measure, the same constructs for both Indigenous and non-Indigenous students. In other words, a large majority of the items seem appropriate for all students in view of the purpose of the tests which is to assess literacy and numeracy knowledge and skills at the start of formal schooling, and are not biased for or against different groups of students in Western Australia. Only a very few items (3% of all items in both scales) showed differential functioning for Indigenous and non-Indigenous students. As far as we are aware, ours is the first study to use the Rasch measurement model to establish this fact. The PIPS-BLA, with the exception of very few items, can thus be accepted as providing fair, valid measures of students' numeracy and literacy skills at the beginning of formal schooling, which may be legitimately compared across groups based on gender, Indigenous or ESL status, and geolocation. Our conclusion using a large sample of Indigenous students is similar to that in a study using a classical approach to assessment and a smaller sample of students (Godfrey & Galloway, 2004).

Once the few DIF and misfitting items were deleted from the analyses, the expectation that Indigenous students, on average, might perform better on these scales than they have done on other scales has not been realised: the same levels of difference in mean scores as were found, for example, in the 2012 NAPLAN Year 3 results, are apparent in the PIPS-BLA results. These sorts of differences have also been reported in studies using other measures and other age groups (for example, see Australian Early Development National Report, 2012; de Bortoli & Thompson, 2010; Bradley, et al, 2007; MCEECDYA, 2010). Our hypothesis that Indigenous students may perform better on average on PIPS because it is individually administered, uses an engaging, pictorial format, and provide measures of students' actual performances rather than teacher ratings of performance, proved not to be upheld. However, although the overall patterns of results are similar to those obtained on other measures, we have no evidence of whether PIPS can provide more valid and reliable, or better predictive measures for any particular student. In addition the range of Indigenous performance is large on both scales with many students performing well on the scales.

The addition of five new items in each scale—items which were expected to be easier than any in the original set of items—in order to better measure low-scoring students, has not been as successful as hoped, but there was some relatively minor improvement in the numbers of students falling below the easiest and second easiest items in both the Reading and Mathematics scales. In order to understand the difficulties encountered by these low-scoring students when undertaking the PIPS-BLA, a study involving teachers' feedback on the test-taking and class-learning behaviours of these students would be helpful. Feedback on a small number of students has indicated that such students have, in the views of their teachers, difficulty understanding the test instructions. Another possibility to pursue might be the students' inability (at this stage in their schooling) to form correspondences between spoken words and pictorial representations of those words, a skill which is necessary to use in both scales.

Indigenous students in Remote communities perform, on average, at a lower level than those Indigenous students in Metropolitan and Regional areas. This is an important point of relevance for classroom practice and parenting classes. Once again, our results support previous findings that Indigenous students in Remote areas are the most at-risk (Australian Early Development Index National Report, 2012; Bradley et al, 2007; Leach, 1999; WAACHS, 2006). The difference in performance on Reading in regard to boys and girls again supports other research, with girls performing better, on average, than boys, but this difference (and in the same direction) is also apparent in performances on Mathematics. And for both scales, girls' performances tend to be less varied than those of boys, so that there tend to be more boys than girls amongst both the lowest-scoring and highest-scoring groups. The differences in Reading in favour of non-ESL students were as expected and are in line with the findings of other studies (Australian Early Development Index National Report, 2012; MCEECDYA, 2010; WAACHS, 2006), but the same difference in regard to Mathematics was less so—however, several aspects of Mathematics which are assessed in PIPS-BLA involve quite a bit of verbal knowledge and hence may explain the similar patterning of mean performance. The mismatch between home and school language and its effect on performance is an important point made by Warren, Young and de Vries (2008), a point which is supported by the current study. In Australia, some Indigenous students' performance would be impacted both by their having English as a second language and living in remote areas.

PIPS-BLA results in their raw form are available to teachers immediately following the completion of the online assessment. In their processed form the data are available within a period of 48 hours. Therefore teachers have ready access to information about each student's literacy, numeracy and phonological awareness, from the start of the school year. Teachers are provided with a range of strategies for developing their skills to interpret their data: workshops; short videos; help desk; phone contact. In all interactions, teachers are supported to develop confidence in interpreting the assessment feedback about their own students. Using the data, they plan appropriate learning programs, building from the unique starting point of each student. This support is particularly useful for teachers – many of whom are novices – located in remote schools, where vast distances make access to face-to-face professional support difficult.

The trend (not statistically significant) in Reading means towards improvement from 2011 to 2012/2013 (a trend which is less marked in Mathematics) may be explained by the fact that the composition of the types and locations of schools which have chosen to undertake the PIPS-BLA testing program has altered from about 2010 forwards, since schools now have the option of being part of a government-sponsored testing program. It may also be that Indigenous students' performance is improving a little, on average. This possibility—which would be a heartening trend—will be monitored in future test administrations.

Acknowledgements
Special thanks to the teachers who administered the PIPS-BLA to their students, to those who provided written feedback, and to the students themselves.

References

Andrich, D. (1988). *Rasch models for measurement.* California: Sage publications.

Andrich, D., Sheridan, B. & Luo, G. (2008). *Rasch models for measurement: RUMM2030.* RUMM Laboratory, Perth, Western Australia.

Andrich, D. & Styles, I. (2004). *Final report on the psychometric analysis of the Early Childhood Development Instrument using the Rasch model: a technical paper commissioned for the development of the Australian early Childhood Development Index (AEDI).* Australia: Royal Children's Hospital.

Australian Commonwealth Government. (1990). *National Policy on Aboriginal Education.* Canberra: Commonwealth Government.

Australian Early Development Index Report (2012). *A snapshot of Early Childhood Development in Australia 2012.* Australian Early Development Index National report. www.rch.org.au/aedi/Report_NationalReport_2012_1304/

Bradley, S., Draca, M., Green, C. & Reeves, G. (2007). The magnitude of educational disadvantage in Indigenous minority groups in Australia. *Journal of Population Economics, 20, 547-569.*

Brookes-Gunn, J. & Duncan, G.J. (1997). The effects of poverty on children. *Future Child, 7, 55-71.*
Catholic Education Commission (2000). *Aboriginal Education Operations Plan.* Perth, Western Australia: Policy statement on Aboriginal Education, Catholic Education Commission. URL: http://web1.ceo.wa.edu.au/pls/portal30/docs/FOLDER/CA CATH ED/CEO REP COMMISSION/POLICIES/SCHOOLANDCOMMUNITYOPERATIONS/2-C1%ABORIGINAL%EDUCATION.pdf

De Bortoli, L. & Thompson, S. (2010). *Contextual factors that influence the achievement of Australia's Indigenous students: Results for PISA 200-2006.* Melbourne: ACER Press.

Early Years Consultative Committee. (2005). *Performance indicators in primary school: Review and evaluation.* Report prepared for the Department of Education, Australian Capital Territory.

Godfrey, J. R. (2003). Report on the administration and analysis of *Performance Indicators in Primary Schools* test to assess literacy skills among Australian Indigenous Children. Round Table Discussion, CEM, University of Durham.

Godfrey, J.R. & Galloway, A. (2004). Assessing early literacy and numeracy skills amongst Indigenous children with the Performance Indicators in Primary Schools test. *Issues in Educational Research, 14.*

Harquist, C. & Andrich., D. (2004). Is The sense of coherence instrument applicable on adolescents? A latent trait analysis using Rasch modeling. *Personality and individual differences, 36, 955-968.*

Harslett, M. (1996). Concept of giftedness from an aboriginal cultural perspective. *Gifted Education International, 11, 100-106.*

Klenowski, V. (2009). Australian Indigenous students: Addressing equity issues in assessment. *Teaching Education, 20, 1, 77-93.*

Leach, A.J. (1999). Otitis media in Australian aboriginal children: An overview. *International Journal of Pediatric Otorhinolaryngology, 49, 1, 173-178.*

Lokan, J., Ford, P., & Greenwood, L. (1997). *Mathematics and Science on the line: Australian middle primary students' performance in the Third International Mathematics and Science Study.* Melbourne: ACER.

Luke, A., Woods, A., Land, R., Bahr, M., & McFarland, M. (2002). *Accountability: Inclusive assessment, monitoring and reporting.* Research report for the Indigenous Education Consultative Study: Brisbane, Queensland.

MCEECDYA. (2010). *Aboriginal and Torres Strait Islander Action Plan 2010-2014.* Ministerial Council for Education, Early Childhood Development and Youth Affairs.

Ministerial Council on Education, Employment, Training and Youth Affairs (1995). *National Strategy for the Education of Aboriginal and Torres Strait Islander Peoples: 1996-2002.* Canberra: Department of Employment, Education, Training and Youth Affairs.

McInerney, D.M. (1991). Key determinants of motivation of non-traditional aboriginal students in school settings: Recommendations for change. *Australian Journal of Education, 35, 2, 154-174.*

Moore, T. (2006). *Wellbeing of Australian children.* Melbourne University Press.

NAPLAN. (2012). http:www.nap.edu.au/verve_resources/naplan_2012_national_report.pdf
Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: MESA Press.

Silburn, S., Brinkman, S., Ferguson Hill, S., Styles, I., Walker, R. & Shepherd, C. (2009). *The Australian Early Development Index (AEDI) Indigenous adaptation study.* Perth, Australia: Curtin University of technology and the Telethon Institute for Child Health Research.
Stobart, G. (2005). Fairness in multicultural assessment. *Assessment in Education, 12, 3, 275-287.*

Styles, I. (2009). *Correlations between matched PIPS-BLA and WALNA data: 2004-2007.* Unpublished report for Pearson Measurement Laboratory, Graduate School of Education, The University of Western Australia.

Taylor, S.D. (1998). *Minority students and gifted and talented programs: perceptions, attitudes and awareness.* Unpublished doctoral thesis: University of Sydney.

Telethon Institute for Child Health Research. (2006*). Improving the educational experiences of Aboriginal Children.* WAACHS Child Health Survey (Vol 3): Perth, Western Australia: Telethon Institute for Child Health Research.

Tripony, P. (2002). *Challenges and tensions in implementing current directions in Indigenous education.* Paper presented at AARE conference, Queensland.

Tymms, P. (1999). *Baseline assessment and monitoring in primary schools' Achievements, attitudes and value-added indicators.* London: David Fulton Publishers.

Tymms, P, Merrell, C., Henderson, B., Albone, S., & Jones, P. (2007). *Links between children's starting points and finishing points in Primary School.* Paper presented at the EARLI Conference, Budapest, 2007.

Wadsworth, M. (1999). Early life. In Marnot, M. &Wilkinson, R.G. (Eds). *Social determinants of health.* Oxford, Oxford University Press, 44-68.

Warren, E., Young, J., & de Vries, E. (2008). *Australian Indigenous Students: The role of oral language and representation in the negotiation of mathematical understandings.* In Watson, J. & K. Beswick. (Eds). Vol 2, Proceedings of the 30th Annual Conference of Mathematics Education Research Group of Australia, 775-884. Mersa: Hobart, Tasmania.

Wildy, H. & Styles, I. (2008a). Measuring what Australian Students entering Primary School know and can do. *Australian Research in Early Childhood Education, 15, 2, 75-85.*

Wildy H. & Styles, I. (2008b). *Psychometric characteristics of the PIPS-BLA Phonological Awareness Scale.* Unpublished report. Graduate School of Education, The University of Western Australia, Crawley, Western Australia.

Authors

Helen Wildy is Dean of the Faculty of Education at The University of Western Australia and since 2001 has directed the PIPS-BLA project in Australia. In her teaching, supervision and research she specialises in the interpretation and use of large scale assessment data to inform decision-making of school leaders and teachers.
Email: helen.wildy@uwa.edu.au

Irene Styles is a research consultant in the Pearson Psychometric Laboratory and a specialist in the development of questionnaires and the use of the Rasch measurement model for data analysis in a wide variety of fields including health sciences, psychology and education. She is also involved in the supervision of postgraduate students.
Email: irene.styles@uwa.edu.au

Vivienne Pepper has been Project Manager at PIPS Australia for over 7 years, both at Murdoch University and The University of Western Australia. She graduated from Curtin University with a Bachelor of Science, and brought Project Management experience with her to the role of PIPS Project Manager.
Email: vivienne.pepper@uwa.edu.au

Joanne Faulkner has managed several research projects at the Graduate School of Education, specializing in assessment of students from Years K – 10. She currently works in the field of organizational and process improvement at Murdoch University.
Email: j.faulkner@murdoch.edu.au

Ye'elah Berman completed her Honours in Psychology at The University of Western Australia and is now a Project Officer in its Faculty of Education.
Email: yeelah.berman@uwa.edu.au