Content validity to support the use of a computer-based phonological awareness screening and monitoring assessment (Com-PASMA) in the classroom

KARYN CARSON1, THERESE BOUSTEAD2 & GAIL GILLON2

1Flinders University, Adelaide, South Australia, and 2University of Canterbury, Christchurch, New Zealand

Abstract

Purpose. This study investigated the content validity of a computer-based phonological awareness (PA) screening and monitoring assessment (Com-PASMA) designed to evaluate school-entry PA abilities. Establishing content validity by confirming that test items suitably ‘fit’ and sample a spectrum of difficulty levels is critical for ensuring educators can deduce accurate information to comprehensively differentiate curricular reading instruction.

Method. Ninety-five children, inclusive of 21 children with spoken language impairment, participated in a 1-year longitudinal study whereby the Com-PASMA was administered at the start, middle and end of the school year.

Result. Estimates of content validity using Rasch Model analysis demonstrated that: (1) rhyme oddity and initial phoneme identity tasks were most appropriate at school-entry and sampled a spectrum of difficulty levels, (2) more challenging phoneme level tasks (e.g. final phoneme identity, phoneme blending, phoneme deletion and phoneme segmentation) became increasingly appropriate and differentiated between high- and low-ability students by the middle and end of the first year of school and (3) letter-knowledge tasks were appropriate but declined in their ability to differentiate student ability as the year progressed.

Conclusion. Findings demonstrate that the Com-PASMA has sufficient content validity to measure and differentiate between the PA abilities of 5-year-old children on entry to school.

Keywords: Phonological awareness, classroom assessment, reading development, content validity, computer-based assessment
**The role of technology in supporting classroom reading practices**

Over the past decade, advances in technology coupled with improved affordability and enhanced utility have seen classrooms worldwide transform into highly connected digitalized environments. In New Zealand, 92% of primary schools report that more than three-quarters of their classrooms are linked to a network (Johnson, Hedditch, & Yin, 2011). Up to three-quarters of New Zealand primary school Principals report that the Internet has a significant impact on teaching and learning practices in their classrooms (Johnson et al., 2011). Notably, the literacy and numeracy initiatives of more than 90% of New Zealand primary schools are supported by resources located online (Johnson et al., 2011). In Australia, the government reports that all schools provide computer and internet access (Australian Government, 2014). Indeed, the Digital Education Revolution, an Australian commonwealth government initiative, supported high-speed broadband internet connectivity to schools, with the states and territories committing to provide enhanced technological infrastructure and improved availability of digitalized resources to schools (Australian Government, 2008). Despite the exponential growth of technology in the classroom, few studies have evaluated the role of technology in supporting educators to differentiate between children with high and low phonological awareness (PA) abilities on entry to schooling.

**Classroom computer-based phonological awareness screening and monitoring**

Although a number of skills support the acquisition of reading proficiency, the omnipresence of technology as a critical contributor to early reading success is widely recognized (Ehri, Nunes, Willows, Schuster, Yaghoub-Zadeh, & Shanahan, 2001). Children who enter school with an impoverished awareness of the syllable, onset-rime and phonemic units of sounds within spoken words are at far greater risk of falling up to 3 years behind in reading acquisition by age 10 compared to their peers who approach reading instruction with these skills (Torgesen, Wagner, & Rashotte, 1994). What is less understood is the role of technology in supporting the routine screening and monitoring of PA as children enter school and engage in their first year of formal education. Few technology-based tools exist to support teachers in measuring the PA skills of children in their classrooms, with even fewer supporting measurement at the critical level of PA – the phoneme level (see Carson, Boustead, & Gillon, 2013a for a review). Many web-based applications that support screening and progress monitoring of literacy skills take the form of a data-management system whereby the educator administers an assessment and then, once completed, enters data into an online system. Examples include the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) (Good & Kaminski, 2003) and the Academic Improvement Measurement System (AIMSweb) (AIMSweb, 2014). Computer-based assessments such as the Cognitive Profiling System (CoPS) (Singleton, Thomas, & Leedale, 1996) and Performance Indicators in Primary School (PIPS) (Tymms, 1999) have the added feature of test administration in addition to automated scoring and storage of results; however, these assessment tools only measure the syllable and onset-rime layers of PA, not the critical phoneme level.

In recognition of the lack of technology-based assessments that measure PA at the phoneme level and can administer, record and score results for classroom educators, the authors developed the freely available, “Computer-based PA Screening and Monitoring Assessment” (Com-PASMA) in 2010. Specifically, development of the Com-PASMA aimed to provide classroom educators with a method of: (a) time-efficient screening and monitoring of skills known to be predictive of reading outcomes; (b) identifying risk for reading impairment from the outset of formal schooling; and (c) differentiating between high- and low-ability students to support curricular decision-making and, in turn, the minimization of large gaps in reading outcomes.

The Com-PASMA is comprised of six PA and two letter-knowledge (LK) tasks and presents all test items, automatically scores responses and stores data into a centralized database. Collectively these features enable children to self-administer the assessment in the classroom. Specific skills measured include rhyme oddity (e.g. which word does not rhyme: cat, mat, bus), initial phoneme identity (e.g. what word starts with the /s/ sound: bee, sun, tent), final phoneme identity (e.g. what word ends with the /t/ sound: hat, hole, sun), phoneme blending (e.g. what word do you think I am saying? m-ou-se), phoneme deletion (e.g. if I say “spin” without the /s/ sound, what word do we get?), phoneme segmentation (e.g. how many sounds do you hear in the word “hand”?), letter-name recognition (e.g. show me the letter “m”) and letter-sound recognition (what letter makes the /mmm/ sound?).

Recent studies show that the Com-PASMA is 30% faster in administration time compared to paper-based equivalents (Carson, Gillon, & Boustead, 2011), produces congruent scores with identical paper-based testing methods (Carson et al., 2011) and can predict reading outcomes at 6-years of age with 94% accuracy when administered at 5 and 5.5 years of age (Carson et al., 2013a). Alongside the reporting of time-efficiency and predictive validity, it is important that content validity be established to ensure test items appropriately “fit” and sample a range of abilities. This is particularly important given that technology-assisted assessment practices may support educators in differentiating between children with high and low PA abilities on
entry to school in an effort to help re-dress international issues related to widening gaps in reading achievement.

**Establishing content validity of computer-based phonological awareness screening and monitoring**

According to Guernsey, Levine, Chiong, and Severns (2012), the technological boom has resulted in a significant quantity of easy-to-download literacy-based applications claiming to support reading development with little evidence of validity, reliability, efficacy or effectiveness to support their claims. The purpose of the current investigation is to report the content validity of the Com-PASMA. Content validity refers to the process of ensuring that test content accurately reflects the knowledge being measured (Anastasi & Urbina, 1997). According to Lissitz and Samuelsen (2007), test items are the building blocks of any measurement tool and are, therefore, a critical source of content validity. Construction of a test with a sufficient degree of content validity includes analysing items within a test to determine whether they are appropriate for the intended population and whether they sample a range of difficulty to enable differentiation between high- and low-ability students (Crocker & Algina, 1986); a particularly important feature for classroom assessment, planning and the minimization of risk for future reading impairment.

A number of statistical methods can be applied to establish the content validity of an assessment tool; one of which is Rasch Model analysis (Bond & Fox, 2007). The Rasch Model is a measurement model whereby the “fit” and difficulty spectrum of test items to an intended population can be formally reviewed and was selected as the model of choice in the current investigation. This model provides a theoretical range against which test developers can compare patterns of responses to determine whether items show a “fit” or “misfit” to the ability of test takers. Items deviating from the ideal range (i.e. showing a misfit) require adaptation or removal from the measurement tool. The model evaluates a test item in terms of difficulty (the item parameter) and people in terms of their ability (the person parameter) to score a correct or incorrect response on a particular test item. The Rasch Model is occasionally referred to as the One-Parameter Logistic Model under Item Response Theory (IRT). Despite sharing a mathematical similarity, these models differ on a conceptual level (Baker, 2001). In the Rasch Model, data must conform to the properties of the model for measurement to take place (Andrich, 2004). Items that do not conform to the model (i.e. show a model “misfit”) require careful investigation and an explanation as to why this is the case. In IRT, the importance of data-model fit is emphasized. However, additional model parameters enable the model to adjust to reflect the pattern of the data (Embretson & Reise, 2000).

To ensure the Com-PASMA contains satisfactory levels of content validity to support educators in accurately differentiating between children with high and low PA abilities, the following hypotheses were proposed:

1. Test items within each task in the Com-PASMA will demonstrate an appropriate “fit” to the abilities of 5-year-old children in the first year of formal schooling.
2. Test items within each task in the Com-PASMA will sample a range of difficulty levels among 5-year-old children.

**Method**

**Participants**

The participants were 95 New Zealand children (39 boys, 56 girls) commencing their first year of formal schooling and were aged between 5 years 0 months and 5 years 2 months ($M = 60.41$, $SD = 0.59$). Of this group, 74 children presented with typical spoken language skills (TD) and 21 children presented with spoken language impairment (SLI). Inclusion of children with (i.e. SLI) and without risk (i.e. TD) for reading problems purposely enabled the authors to evaluate whether test items in each task of the Com-PASMA demonstrated an appropriate “fit” to the ability spectrum of the intended population and sampled a wide spectrum of difficulty. All participants presented with typical cognitive, hearing and physical development and attended a mainstream school. The participants formulated a control sample for another larger study evaluating the effectiveness of teacher-directed classroom PA instruction (see Carson, Gillon, & Boustead, 2013b).

**Classification of SLI and TD**

To be classified as having TD, children needed to score within or above the average range on the following assessments: (a) Clinical Evaluations of Language Fundamentals—Preschool, Second Edition ( CELF-P2) (Wiig, Secord, & Semel, 2006) receptive and expressive language indices, where a composite score of 85–115 is considered the average range; (b) Preschool and Primary Inventory of PA (PIPA) (Dodd, Crosbie, McIntosh, Teitzel, & Ozanne, 2000) rhyme oddity, initial phoneme identity and letter-sound knowledge sub-tests, where a standard score of 7–13 is considered the average range; (c) Primary Test of Non-verbal Intelligence (PTONI) (Ehrler & McGhee, 2008), where a score of 85–115 is considered the average range; and (d) New Zealand Articulation Test (NZAT) (Ministry of Education, 2004) single consonants and consonant blends sub-tests, where a PCC above 93% for 5-year-old children is considered satisfactory (Shriberg, Austin, Lewis, McSweeny, & Wilson, 1997).
To be classified as having SLI, children needed to perform at least one standard deviation below the mean on one of the baseline language measures (e.g. CELF-P2 or PIPA) or present with phonologically based speech errors with a PCC below 93%. In practice, children who perform at least 1–2 SD below the mean are considered to have impaired language skills and, thus, a cut-off of at least 1 SD below the mean was used in this study. A PCC of 93.4–100% is considered typical for children aged 5 years and 0 months to 5 years and 11 months (Shriberg et al., 1997). Children in this study who had a PCC below 93% were considered to present with phonologically based speech impairment, ranging in severity from mild-to-severe. The majority of children (81%) classified as having SLI presented with deficits in both language and speech development. Table I profiles the language, PA, speech and non-verbal intellectual abilities of children with SLI and TD at the start of the study.

### Procedure

A 12-month longitudinal research design was employed whereby each participant completed all tasks in the Com-PASMA at the start, middle and end of the first year at school. Test items within each task were systematically reviewed using Rasch Model analysis to establish content validity of the Com-PASMA at each of the three points throughout the school year.

### Assessment measures

For the purposes of the current investigation, test items in the Com-PASMA are the assessment measures of focus. Each participant completed each task in the Com-PASMA at ~ 5 (i.e. start of the year), 5.5 (i.e. middle of the year) and 6-years of age (i.e. end of the year). Children were assessed individually under the supervision of a qualified speech-language pathologist in a quiet area of the school environment. The Com-PASMA contains six PA and two LK tasks (i.e. rhyme oddity, initial phoneme identity, final phoneme identity, phoneme blending, phoneme deletion, phoneme segmentation, letter-name recognition, letter-sound recognition). Tasks take ~ 4–5 minutes to complete. Calculations of sensitivity and specificity from the start and middle of the school year with reading outcomes at the end of the school year are 0.89 and 0.95, respectively (Carson et al., 2013a). Average performances on each task in the Com-PASMA at start, middle and end of the school year are profiled in Table II for the entire study sample and for the sub-set of children with SLI.

### Scoring reliability

Tasks in the Com-PASMA were administered individually during one session at the start, middle and end of the school year. Data was recorded in real-time as well as by the automated scoring system of the Com-PASMA. In addition, all assessment sessions were recorded using a DVD recorder. Half of the assessment sessions were viewed anonymously and in their entirety by an independent researcher at the start, middle and end of the school year. One-hundred per cent inter-rater reliability agreement between real-time scoring, automated Com-PASMA scoring and DVD scoring was achieved for all tasks and test items that were reviewed.

| Table I. Spoken language profiles of children with TD and SLI at the start of the school year. |
|-----------------|-----------------|
|                | TD (n = 74)     | SLI (n = 21)   |
|                | M               | SD              | Range          | M               | SD              | Range          |
| Age            | 60.34 mths      | 0.58            | 60–70 mths     | 60.42 mths      | 0.60            | 60–70 mths     |
| CELF-P2        |                 |                 |                |                 |                 |                |
| Receptive Language | 102.13         | 8.95            | 86–119         | 85.29**         | 4.00            | 79–95          |
| Expressive Language | 100.32         | 8.88            | 86–116         | 84.71**         | 3.48            | 82–93          |
| PIPA           |                 |                 |                |                 |                 |                |
| Rhyme Oddity   | 6.96            | 2.33            | 3–13           | 4.10**          | 1.58            | 2–7            |
| Initial Phoneme Identity | 9.84     | 2.44            | 3–15           | 6.76**          | 1.30            | 6–10           |
| Letter-Sound Recognition | 7.8     | 1.86            | 4–11           | 5.05**          | 1.12            | 4–7            |
| PTONI          | 105.59          | 5.69            | 87–116         | 94.43**         | 6.30            | 85–103         |
| PCC            | 97.59           | 5.56            | 94–100         | 77.86%**        | 11.83           | 65–98          |

CELF-P2 Receptive Language, Receptive Language Index of the Clinical Evaluations of Language Fundamentals – Preschool, Second Edition (M = 100, SD = ±15) (Wiig et al., 2006); CELF-P2 Expressive Language, Expressive Language Index of the Clinical Evaluations of Language Fundamentals – Preschool, Second Edition (M = 100, SD = ±15) (Wiig et al., 2006); PIPA, Preschool and Primary Inventory of PA (M = 10, SD = ±3) (Dodd et al., 2000); PTONI, Primary Test of Non-verbal Intelligence (M = 100, SD = ±15) (Ehri & McGhee, 2008); TD, typical development; SLI, spoken language impairment; mths, months; PCC, percentage of consonants spoken correctly on the New Zealand Articulation Test.

**Children with SLI performed significantly lower than children with TD (p < 0.001).
showed a significant model “fit” or “misfit” and provided information on: (a) which test items steps (Version 3.70) (Linacre, 2010). This analysis responses into a software programme called Win-

Analysis procedure

Rasch Model analysis was conducted on the responses of the 95 children to each test item at the start, middle and end of the school year by entering responses into a software programme called Winsteps (Version 3.70) (Linacre, 2010). This analysis provided information on: (a) which test items showed a significant model “fit” or “misfit” and (b) how the test items in each task related to each other in terms of difficulty. Using Winsteps “simulate data” option, a minimum sample of 64 participants was required to achieve a confidence interval of 95% with a ± ½ logit score.

Winsteps provides several types of statistical analyses to evaluate test items in relation to the latent trait being measured. For the purposes of our analysis, the “outfit statistics” of mean-square and ZSTD for each test item were used to evaluate which items showed a “fit” or significant model “misfit”. The relevance of using the mean-square statistic and the ZSTD statistic are described below:

- **Mean-square statistic (MNSQ):** The mean-square statistic draws attention to the accuracy of an item by providing an indication of the size of an item’s “misfit” to the model. An item with a mean square close to 1.0 suggests that the item is accurate. An item with a mean square less than 1.0 is considered less accurate, but this does not cause any real problems (Linacre, 2010, p. 23). An item with a mean square greater than 2.0 is considered inaccurate and in need of attention.

- **ZSTD statistic:** A ZSTD statistic is assigned to each mean-square statistic to indicate whether the size of the “misfit” is statistically significant. The ZSTD is “standardized like a Z-score” (Linacre, 2010, p. 25). An item with a ZSTD statistic between -2 and +2 indicates a statistically significant model “fit”.

In line with Linacre (2010), test items with a mean-square statistic greater than 2.0 and a ZSTD statistic less than -2 and greater than +2 were interpreted as showing a statistically significant model “misfit”. Items may show a “misfit” for a number of reasons, including: (1) being too easy or difficult, (2) confusing or ambiguous instructions or (3) lack of image quality (i.e. animated or static graphics) or heightened linguistic complexity. Items demonstrating a “misfit” may require adaption or deletion from the instrument (Bond & Fox, 2007).

The point-measure correlation is another outfit statistic. It provides an indication of an item’s discrimination ability. Positive correlations above 0.3 indicate that the item is well correlated to the ability being measured. Negative correlations or correlations close to zero suggest there is little relationship between the item and the ability being measured. This indicates that an item does not effectively distinguish between individuals with more or less ability, which is cause for concern.

Results

Using Rasch Model analysis, Tables III and IV profile examples of the mean-square statistic, ZSTD statistic and point-measure correlation for test items at the start of the school year in the rhyme oddity and phoneme segmentation tasks, respectively. Rhyme oddity was considered to be the easiest PA task, as well as appropriate for 5-year-old children. Therefore, it was anticipated that the majority of rhyme oddity test items would demonstrate a model “fit” and point-measure correlation above 0.3. Pho-

Analysis procedure

Rasch Model analysis was conducted on the responses of the 95 children to each test item at the start, middle and end of the school year by entering responses into a software programme called Winsteps (Version 3.70) (Linacre, 2010). This analysis provided information on: (a) which test items showed a significant model “fit” or “misfit” and (b) how the test items in each task related to each other in terms of difficulty. Using Winsteps “simulate data” option, a minimum sample of 64 participants was required to achieve a confidence interval of 95% with a ± ½ logit score.

Winsteps provides several types of statistical analyses to evaluate test items in relation to the latent trait being measured. For the purposes of our analysis, the “outfit statistics” of mean-square and ZSTD for each test item were used to evaluate which items showed a “fit” or significant model “misfit”. The relevance of using the mean-square statistic and the ZSTD statistic are described below:

- **Mean-square statistic (MNSQ):** The mean-square statistic draws attention to the accuracy of an item by providing an indication of the size of an item’s “misfit” to the model. An item with a mean square close to 1.0 suggests that the item is accurate. An item with a mean square less than 1.0 is considered less accurate, but this does not cause any real problems (Linacre, 2010, p. 23). An item with a mean square greater than 2.0 is considered inaccurate and in need of attention.

- **ZSTD statistic:** A ZSTD statistic is assigned to each mean-square statistic to indicate whether the size of the “misfit” is statistically significant. The ZSTD is “standardized like a Z-score” (Linacre, 2010, p. 25). An item with a ZSTD statistic between -2 and +2 indicates a statistically significant model “fit”.

In line with Linacre (2010), test items with a mean-square statistic greater than 2.0 and a ZSTD statistic less than -2 and greater than +2 were interpreted as showing a statistically significant model “misfit”. Items may show a “misfit” for a number of reasons, including: (1) being too easy or difficult, (2) confusing or ambiguous instructions or (3) lack of image quality (i.e. animated or static graphics) or heightened linguistic complexity. Items demonstrating a “misfit” may require adaption or deletion from the instrument (Bond & Fox, 2007).

The point-measure correlation is another outfit statistic. It provides an indication of an item’s discrimination ability. Positive correlations above 0.3 indicate that the item is well correlated to the ability being measured. Negative correlations or correlations close to zero suggest there is little relationship between the item and the ability being measured. This indicates that an item does not effectively distinguish between individuals with more or less ability, which is cause for concern.

Results

Using Rasch Model analysis, Tables III and IV profile examples of the mean-square statistic, ZSTD statistic and point-measure correlation for test items at the start of the school year in the rhyme oddity and phoneme segmentation tasks, respectively. Rhyme oddity was considered to be the easiest PA task, as well as appropriate for 5-year-old children. Therefore, it was anticipated that the majority of rhyme oddity test items would demonstrate a model “fit” and point-measure correlation above 0.3. Pho-

### Table II. Mean (SD) performance on the Com-PASMA at the start, middle and end of the first year of school for the entire sample (n = 95) and the sub-set of children with SLI (n = 21).

|                      | School-entry (5;0–5;2)* | Middle of year (5;5–5;7)* | End of year (5;11–6;01)* |
|----------------------|-------------------------|---------------------------|--------------------------|
| Phonological awareness** | Entire sample | Sub-set of children with SLI | Entire sample | Sub-set of children with SLI | Entire sample | Sub-set of children with SLI |
| - RO                 | 5.3 (2.5)               | 2.2 (0.8)                 | 7.5 (2.5)               | 5.1 (2.5)                 | 8.6 (1.8)               | 7.3 (1.6)                 |
| - IPI                | 5.3 (3.0)               | 1.5 (0.2)                 | 7.6 (2.6)               | 4.3 (2.9)                 | 8.5 (2.3)               | 5.5 (2.3)                 |
| - FPI                | 0.6 (0.7)               | 0.1 (0.5)                 | 4.8 (2.9)               | 1.2 (1.1)                 | 6.4 (3.1)               | 2.8 (2.6)                 |
| - PB                 | 1.3 (1.2)               | 0.8 (0.3)                 | 8.4 (5.2)               | 4.0 (2.1)                 | 11.7 (3.74)            | 8.0 (3.4)                 |
| - PD                 | 0.5 (0.8)               | 0.2 (0.3)                 | 6.2 (4.2)               | 3.2 (1.3)                 | 8.8 (3.5)               | 5.2 (3.0)                 |
| - PS                 | 0.6 (0.9)               | 0.1 (0.1)                 | 5.9 (4.1)               | 3.1 (0.8)                 | 8.0 (4.0)               | 5.1 (2.9)                 |
| - LN                 | 13.4 (4.6)              | 5.1 (2.0)                 | 15.8 (3.2)              | 10.1 (2.2)                | 16.7 (2.5)              | 12.3 (3.5)                |
| - LS                 | 11.1 (5.4)              | 3.4 (1.5)                 | 14.3 (4.5)              | 8.9 (3.4)                 | 15.9 (3.4)              | 11.4 (2.9)                |

*Age range (years; months); ** Tasks in the Com-PASMA where RO = rhyme oddity, IPI = initial phoneme identity, FPI = final phoneme identity, PB = phoneme deletion, PD = phoneme segmentation, LN = letter-name recognition; PS = letter-sound recognition; numbers in brackets = standard deviations.
Table III. “Fit” or “misfit” for rhyme oddity test items at school-entry.

| Test items | Which word does not rhyme? | Outfit statistics | PT-measure correlation | Interpretation “Fit” or “Misfit” | % of participants responding correctly |
|------------|----------------------------|-------------------|------------------------|----------------------------------|--------------------------------------|
| 1          | cat mat bus                | 5.98 2.81         | 0.29                   | Misfit                           | 87%                                  |
| 2          | peg doll leg               | 2.22 3.01         | 0.29                   | Misfit                           | 50%                                  |
| 3          | sand hand cup              | 1.27 0.72         | 0.66                   | Fit                              | 40%                                  |
| 4          | hen car pen                | 0.63 0.61         | 0.77                   | Fit                              | 59%                                  |
| 5          | dog book hook              | 0.37 0.92         | 0.62                   | Fit                              | 69%                                  |
| 6          | bun sun kite               | 0.82 0.13         | 0.52                   | Fit                              | 30%                                  |
| 7          | tent lock sock             | 1.66 1.53         | 0.62                   | Fit                              | 48%                                  |
| 8          | shell duck bell            | 0.33 1.01         | 0.58                   | Fit                              | 20%                                  |
| 9          |                             | 0.36 0.91         | 0.59                   | Fit                              | 19%                                  |

Table III demonstrates that, of the rhyme oddity test items, eight showed a model “fit” and items 1 and 2 demonstrated a significant model “misfit”. Inspection of responses to item 1 revealed that 87% of children responded correctly to this item, suggesting that the “misfit” occurred because the item was too easy. Inspection of responses to item 2 revealed that this item was of average difficulty and that the “misfit” could not be attributed to the item being too easy or too difficult for 5-year-old children. In addition, the point-measure correlations for items 1 and 2 were just below 0.3, indicating that these items do not differentiate well between high- and low-ability students.

Table IV shows that phoneme segmentation items 1, 2 and 4 demonstrate a model “fit”, while items 3, 5 to 18 due to lack of variance in the dataset may not be appropriate for measuring PA ability in 5-year-old children at school-entry.

Table IV. “Fit” or “misfit” for phoneme segmentation test items at school-entry.

| Test items | How many sounds in the word …? | Outfit statistics | PT-measure correlation | Interpretation “Fit” or “Misfit” | % of participants responding correctly |
|------------|--------------------------------|-------------------|------------------------|----------------------------------|--------------------------------------|
| 1          | moon (3)                       | 0.63 −0.41        | 0.79                   | Fit                              | 70%                                  |
| 2          | tooth (3)                      | 0.56 −0.41        | 0.87                   | Fit                              | 56%                                  |
| 3          | cow (2)                        | 2.08 −2.82        | 0.29                   | Misfit                           | 9%                                   |
| 4          | cup (3)                        | 0.08 −1.41        | 0.45                   | Fit                              | 51%                                  |
| 5          | soap (3)                       |                   |                        | Misfit                           | 8%                                   |
| 6          | saw (2)                        |                   |                        | Misfit                           | 8%                                   |
| 7          | flush (4)                      | 4% 8              |                        | Misfit                           | 5%                                   |
| 8          | crab (4)                       |                   |                        | Misfit                           | 4%                                   |
| 9          | sew (2)                        |                   |                        | Misfit                           | 3%                                   |
| 10         | step (4)                       |                   |                        | Misfit                           | 3%                                   |
| 11         | plate (4)                      |                   |                        | Misfit                           | 3%                                   |
| 12         | star (3)                       |                   |                        | Misfit                           | 2%                                   |
| 13         | bank (4)                       |                   |                        | Misfit                           | 1%                                   |
| 14         | lock (3)                       |                   |                        | Misfit                           | 1%                                   |
| 15         | jump (4)                       |                   |                        | Misfit                           | 1%                                   |
| 16         | pond (4)                       |                   |                        | Misfit                           | 2%                                   |
| 17         | bear (2)                       |                   |                        | Misfit                           | 1%                                   |
| 18         | cast (4)                       |                   |                        | Misfit                           | 1%                                   |

3/18 items = Fit
15/18 items = Misfit

(number) = the number in brackets indicates the number of phonemes in the target word according to New Zealand English pronunciation; MNSQ, mean-square statistic; ZSTD, ZSTD statistic; PT-measure correlation, point-measure correlation; “Fit” or “Misfit” indicates whether the data fit the properties of the Rasch Model; italics = items that show a significant model “misfit” and may not be appropriate for measuring PA ability in 5-year-old children at school-entry.
5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17 and 18 demonstrate a significant model “misfit”. Items 1 (“moon”), 2 (“tooth”) and 4 (“cup”) were simple CVC words containing either long medial vowels, earlier developing sounds (e.g. /k/, /p/, /m/) or both. This may have increased the salience of the phonemes in these items, allowing 5-year-old children to segment them correctly at school-entry. Item 3 (“cow”) was thought to be easy during the construction of the test. However, over 85% of children indicated that this item had three sounds, as opposed to two. From video observations, it was revealed that children often added a final schwa phoneme when segmenting “cow” verbally (i.e. /kaʊə/). Only 9% of children scored correctly on this item. The majority of phoneme segmentation items demonstrating a significant “misfit” had a CCVC or CVCC syllable structure. Less than 10% of participants provided a correct response to phoneme segmentation items because they were extremely difficult.

Table V aggregates the mean-square statistic, ZSTD statistic and point-measure correlation for all test items at the start, middle and end of the first year at school. This table illustrates that 108 out of 114 test items demonstrate a model “fit” at one or multiple points throughout the first year at school. All items demonstrating a model “fit” also had point-measure correlations above 0.3, indicating that they discriminate well between individuals of high- and low-ability. Items demonstrating a model “misfit” may do so because of test item difficulty (as described in the following section) or another extraneous factor such as instruction quality, as opposed to the presence of SLI.

### Identifying a hierarchy of item difficulty in each Com-PASMA task

Rasch analysis enables comparison of item difficulty through computation of “estimates of item difficulty” (Bond & Fox, 2007). Winsteps refers to “estimate of item difficulty” as the “measure” statistic, which in essence is a logit (log-odds) score assigned to an item to indicate its difficulty (Linacre, 2010). A logit score is plotted along an interval scale called a logit scale. The logit value of zero represents an arbitrary mean. Therefore, items with a logit score near zero are considered to be of average difficulty. Items with increasingly positive logit scores are more difficult, while items with increasingly negative logit scores are less difficult. The logit value of zero is the threshold for distinguishing between correct and incorrect responses; a logit score above zero suggests correct responses, and a logit score below zero suggests incorrect responses.

### Table V. Summary of items by task demonstrating a “fit” or significant model “misfit” at the start, middle and end of the school year.

| PA task          | Start of year (Age: 5:0–5:2) | Middle of year (Age: 5:5–5:7) | End of year (Age: 5:11–6:01) |
|------------------|--------------------------------|--------------------------------|-------------------------------|
|                  | Fit                            | Significant misfit              | Fit                           | Significant misfit              | Fit                           | Significant misfit              |
| Rhyme Oddity     | 1 (easy), 2                    |                                 | 1 (easy), 2                   |                                 | (easy), 2                      |                                 |
| Initial Phoneme  | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  |                                 | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 |                                 | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  |                                 |
| Identity         |                                |                                 |                               |                                 |                               |                                 |
| Final Phoneme    | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  | 2 (hard)                        | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  |                                 | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  | (easy), 2                      |
| Identity         | 9, 10                          |                                 | 9, 10                         |                                 | 9, 10                         |                                 |
| Phoneme          | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  |                                 | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  |                                 | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  |                                 |
| Blending         |                                |                                 |                               |                                 |                               |                                 |
| Phoneme          | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  | 2 (hard)                        | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  |                                 | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  | (easy), 2                      |
| Deletion         | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  | 2 (hard)                        | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  |                                 | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  | (easy), 2                      |
| Phoneme          | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  | 2 (hard)                        | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  |                                 | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  | (easy), 2                      |
| Segmentation     | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  | 2 (hard)                        | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  |                                 | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  | (easy), 2                      |
| Letter-name      | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  | 2 (hard)                        | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  |                                 | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  | (easy), 2                      |
| Letter-sound     | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  | 2 (hard)                        | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  |                                 | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  | (easy), 2                      |

Age is represented as years;months; numerals in the table indicate the test item “number”; (easy) indicates an item on which the majority of participants responded correctly, suggesting the “misfit” occurred because the item was too easy; (hard) indicates an item on which the majority of participants responded incorrectly, suggesting the “misfit” occurred because the item was too hard; items showing a “misfit” without an “easy” or “hard” description show a “misfit” for a reason besides difficulty level and require further investigation (i.e. misfit may be due to variables such as linguistic complexity or image quality).
items with increasingly negative logit scores are easier. In theory, a logit scale can range from negative infinity to positive infinity (Bond & Fox, 2007). Therefore, for the purposes of this investigation, the following difficulty descriptions were applied to logit values: 8 and above = very difficult; 5–7 = difficult; 2–4 = moderately difficult; 1 to -1 = average difficulty; -2 to -4 = moderately easy; -5 to -7 = easy; and -8 and below = very easy.

Supplementary Tables A, B and C to be found online at http://informahealthcare.com/doi/abs/10.3109/17549507.2015.1016107 summarize the hierarchy of difficulty for test items in each task at the start, middle and end of the school year by plotting the “measure” statistic (i.e. logit score) for each item against a logit scale. From the start to the middle of the school year (i.e. Supplementary Tables A and B to be found online at http://informahealthcare.com/doi/abs/10.3109/17549507.2015.1016107), items in the final phoneme identity, phoneme blending, phoneme deletion and phoneme segmentation tasks begin to sample a wider range of difficulty levels. This is likely because tasks and items that were more difficult at school-entry became easier as children’s PA skills developed.

From the middle to the end of the school year (i.e. Supplementary Tables B and C to be found online at http://informahealthcare.com/doi/abs/10.3109/17549507.2015.1016107), test items became increasingly easier for children to complete as they approached 6-years of age.

Ensuring test item “fit” and difficulty are supported by reliable test parameters

Internal consistency between items within each PA task at the start, middle and end of the first year at school were calculated using Cronbach’s alpha to profile stability at each assessment point as well as across time. Cronbach’s alpha scores above 0.7 indicate that the items within a task are internally consistent (Field, 2009). Table VI profiles the Cronbach’s alpha scores for each task throughout the school year.

At school-entry, rhyme oddity and initial phoneme identity showed a high degree of internal consistency, with Cronbach’s alpha scores of 0.81 and 0.89, respectively. In addition, the letter-name and letter-sound tasks achieved scores of 0.81 and 0.80, respectively. Unsatisfactory Cronbach’s alpha scores were calculated for final phoneme identity, phoneme blending, phoneme deletion and phoneme segmentation at the start of the year. This is consistent with Rasch analysis findings, indicating that these latter tasks are generally more difficult and less reliable at school-entry. By the middle and end of the school year, high Cronbach’s alpha scores were calculated for all tasks in the Com-PASMA.

Test–re-test reliability refers to the consistency of a test across repeated administrations under identical conditions (Thorndike & Thorndike-Christ, 2010). Test–re-test reliability was conducted by correlating each task with itself at each of the three assessment points during the school year. The resulting correlation matrix, shown in Table VII, revealed significant correlations at $p<0.01$ for each task.

Collectively, internal consistency and test–re-test reliability provide strong evidence that the Com-PASMA is internally reliable and highly consistent over repeated administrations.

Discussion

This study investigated the content validity of a computer-based PA screening and monitoring assessment (Com-PASMA). Content validity was established through systematic item review using Rasch Model analysis to determine: (a) appropriateness of test item “fit” to the PA ability spectrum of 5-year-old children and (b) sufficient sampling of a range of difficulty levels to support differentiating of student abilities to inform curricular reading instruction.

Ensuring appropriateness of “fit” to the ability spectrum of 5-year-old children

The first hypothesis predicted that test items in each of the eight tasks in the Com-PASMA would

Table VII. Correlations between each phonological awareness task at the start, middle and end of the school year.

|          | 2 RO | 3 RO | 2 PB | 3 PB |
|----------|------|------|------|------|
| Start of year |      |      |      |      |
| (5.0–5.2) |      |      |      |      |
| Middle of year |      |      |      |      |
| (5.5–5.7) |      |      |      |      |
| End of year |      |      |      |      |
| (5.11–6.01) |      |      |      |      |
| Rhyme oddity | 0.81 | 0.85 | 0.84 |      |
| Initial phoneme identity | 0.89 | 0.85 | 0.85 |      |
| Final phoneme identity | 0.14 | 0.84 | 0.89 |      |
| Phoneme blending | 0.45 | 0.94 | 0.92 |      |
| Phoneme deletion | 0.15 | 0.94 | 0.85 |      |
| Phoneme segmentation | 0.47 | 0.89 | 0.89 |      |
| Letter-name | 0.81 | 0.82 | 0.82 |      |
| Letter-sound | 0.80 | 0.81 | 0.82 |      |

Age is represented as years;months.
demonstrate an appropriate “fit” to the abilities of 5-year-old children in the first year of formal schooling. Rasch Model analysis using the “outfit statistics” of mean-square, ZSTD and point-measure correlation demonstrated that the majority of test items (i.e. 108 out of 114) across the eight tasks showed a model “fit” at one or multiple points during the first year at school. At school-entry, phoneme segmentation was the only task where the majority of test items (15 out of 18 test items) did not show a model “fit”. This suggests that phoneme segmentation items, as part of the Com-PASMA, are less suitable for measuring the PA ability of 5-year-old children at school-entry. However, by the middle and end of the school year, 14 out of 18 phoneme segmentation items demonstrated a “fit”, indicating that these items become increasingly appropriate measures as children begin to interact with beginning classroom literacy instruction. This is consistent with research findings showing that phoneme segmentation is a difficult task at 5 years of age (Adams, 1990). Importantly, Rasch Model analysis demonstrated that, as PA tasks, including phoneme segmentation, became increasingly appropriate across the first year of schooling, the capacity of those items to differentiate between high- and low-ability students increased, as evidenced through improved point-measure correlation statistics.

Of the 114 test items, six items demonstrated a significant model “misfit” at all three assessment points in the school year. These items were rhyme oddity items 1 and 2, phoneme deletion item 1 and phoneme segmentation items 16, 17 and 18. Other items demonstrated a “misfit” at only one or two assessment points in the school year. These include rhyme oddity item 4, final phoneme identity items 2 and 4, phoneme deletion items 6, 5 and 11 and phoneme segmentation item 1. Adapting these “misfit” items in terms of linguistic complexity (i.e. word familiarity, syllable structure and manner of articulation) and presentation (i.e. animated and static graphics or verbal instructions) at the point at which the “misfit” occurred will be required in future investigations. However, some items demonstrating a significant “misfit” may not necessarily require adaption. This is because their low level of difficulty is a purposeful part of test construction, to help ensure graded levels of difficulty within tasks. For example, rhyme oddity item 1 was developed using a simple syllable structure, a high-frequency rhyme unit and salient contrasts in the manner of articulation between correct and distractor options. This was done to ensure children’s success and familiarity with the responding procedure on what would usually be one of the first tasks administered as part of the Com-PASMA.

**Sampling a range of difficulty levels among 5-year-old children**

The second hypothesis stated that test items within each task in the Com-PASMA would sample a range of difficulty levels among 5-year-old children. Rasch Model analysis confirmed this hypothesis. Using “estimates of item difficulty”, results demonstrated that, at school-entry, test items in the Com-PASMA sampled a wide range of difficulty levels (see Supplementary Table A to be found online at http://infor-mahealthcare.com/doi/abs/10.3109/17549507.2015.1016107). Rhyme oddity and initial phoneme identity items provided an even spectrum of easier items to those that are more difficult. For example, rhyme oddity items ranged from moderately easy to moderately difficult and initial phoneme identity items ranged from easy to difficult. Final phoneme identity, phoneme blending and phoneme deletion items predominantly sampled the moderately difficult to difficult range. For example, eight out of 10 final phoneme identity items were classified as difficult, while nine out of 15 phoneme blending and 10 out of 15 phoneme deletion items were considered moderately difficult. This suggests that, while the majority of test items in these tasks demonstrate a “fit” at school-entry, their spectrum of difficulty indicates they may be more appropriate later in the first year of schooling. Only four out of 18 phoneme segmentation items could be analysed at school-entry; 14 items were extremely difficult and could not be analysed because the majority of respondents scored incorrectly. This suggests that these items do not adequately sample a range of difficulty levels at this stage of schooling. The majority of letter-name items were moderately easy, whereas letter-sound items tended to be of average difficulty and are considered appropriate at school-entry. During the middle of the school year (see Supplementary Table B to be found online at http://infor-mahealthcare.com/doi/abs/10.3109/17549507.2015.1016107) rhyme oddity and initial phoneme identity continued to sample a range of easy to moderately difficult ability levels, with initial phoneme identity items being less spread than at school-entry. While the majority of final phoneme identity, phoneme blending, phoneme deletion and phoneme segmentation items continued to be of greater difficulty, they began to sample the moderately easy to easy range. For example, four out of 15 phoneme-blending items, five out of 15 phoneme deletion items and five out of 18 phoneme segmentation items were either moderately easy or easy by the middle of the school year. This is in comparison to one out of 15 phoneme blending items, one out of 15 phoneme deletion items and two out of 18 phoneme segmentation items being classified as moderately easy or easy at school-entry. Letter-name and letter-sound test items became easier to complete by the middle of the school year. By the end of the school year (see Supplementary Table C to be found online at http://infor-mahealthcare.com/doi/abs/10.3109/17549507.2015.1016107), test items were more evenly spread across high to low logit scores, particularly for items in the final phoneme...
identity, phoneme deletion and phoneme segmentation tasks. Interestingly, phoneme deletion item 11, which was moderately difficult at the start and middle of the school year, became very difficult by the end of the school year and will require further investigation in future studies.

Implications for classroom assessment practices

Providing evidence of content validity through the use of Rasch Modal analysis has implications for classroom practices in that educators are informed of which PA tasks are appropriate at exactly which stages of the first year at school. Specifically, aggregation of test item “fit” with “estimates of test item difficulty” demonstrate that: (1) rhyme oddity and initial phoneme identity test items are most appropriate at school-entry and sample a spectrum of difficulty levels, (2) more challenging phoneme-level test items (e.g. final phoneme identity, phoneme blending, phoneme deletion and phoneme segmentation) become increasingly appropriate and differentiate between high- and low-ability students by the middle and end of the first year of school and (3) letter-knowledge test items are appropriate but decline in their ability to differentiate between high- and low-ability students as the first year of schooling progresses. Such findings are consistent with previous research on the developmental progression of increasingly complex PA skills (Elbro, Borstrom, & Petersen, 1998; van Bon & van Leeuwe, 2003). Although the majority of test items in the Com-PASMA demonstrate a “fit” to the ability spectrum of 5-year-old children, sample a range of difficulty levels and discriminate well in the first year of schooling, educators should select those tasks within the Com-PASMA that are best suited to the age-level (i.e. 5-years, 5.5 years and 6-years) of their students when using this tool in the classroom.

Limitations

In light of the positive outcomes of this investigation, it is noteworthy to identify limitations that can be addressed by future research. First, the study consisted of a small sample size of 95 participants, all within the same age range. Future investigations could enhance the current investigation by increasing the sample size and measuring the PA abilities of children both prior to, during and after the first year of schooling to identify where test items are appropriate for a range of age-levels. Second, the Com-PASMA currently consists of a total of 114 test items over eight tasks. Future investigations should ideally focus on constructing a larger range of test items to help support the development of an adaptive version of the assessment. Computer-adaptive testing requires large banks of test items so that the computer is able to select appropriate test items when adapting to the responses of the child. The number of test items in the current investigation that demonstrate a “fit” at some stage in the school year and sample a range of difficulty levels provide a starting point for the expansion of test items and the construction of an adaptive form of the Com-PASMA.

Ensuring all young children have the opportunity to develop proficiency in reading is currently a key area of interest globally. Large gaps in reading outcomes between strong and weaker readers, particularly among developed countries, has created a need for researchers to investigate methods that support the early identification and prevention of risk for reading impairment as part of daily curricular practices. Providing educators with technology-based assessment methods that have demonstrated a robust ability to differentiate between children who enter school with high and low levels of skills, such as PA, that are known to predict later reading success, is perhaps on way in which educators, researchers and policymakers can help reduce the prevalence of large gaps in school-aged reading outcomes.

Acknowledgements

The authors extend thanks to the children, families and educators involved in this study. Gratitude is expressed towards the New Zealand Tertiary Education Commission for financial support in the implementation of this project.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

Adams, M. J. (1990). Beginning to read: Thinking and learning about print. Cambridge, MA: MIT Press.
AIMSweb. (2014). AIMSweb – Always learning. Upper Saddle River, NJ: Pearson. Available online at: http://www.aimsweb.com/.
Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed.). Upper Saddle River, NJ: Prentice Hall.
Andrich, D. (2004). Controversy and the Rasch Model: A characteristic of incompatible paradigms? Medical Care, 42, 1–16.
Australian Government. (2008). Success through partnership: Achieving a national vision for ICT in schools – Strategic plan to guide the implementation of the Digital Education Revolution initiative and related initiatives. Canberra, Australia: Australian Government. Available online at: http://www.teacherstandards.aust.edu.au/static/docs/hotTopics/Success_through_partnership_-_Achieving_a_national_vision_for_ICT_in_schools.pdf.
Australian Government. (2011). Review of funding for schooling – Final report. Canberra, Australia: Author.
Australian Government. (2014). Schools in Australia. Canberra, Australia: Australian Government. Available online at: http://www.studyinaustralia.gov.au/global/australian-education/schools.
Baker, F. (2001). The basics of item response theory: ERIC clearinghouse on assessment and evaluation. University of Maryland, College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.
Supplementary material available online

Supplementary Tables A, B and C to be found online at http://informahealthcare.com/doi/abs/10.3109/17549507.2015.1016107.