Assessment in Education: Principles, Policy & Practice

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/caie20

Judgement-based performance measures of literacy for students with additional needs: seeing students through the eyes of experienced special education teachers

Kerry Woods[a] & Patrick Griffin[a]

[a] Assessment Research Centre, Melbourne Graduate School of Education, University of Melbourne, Parkville, VIC, Australia.

Published online: 22 Oct 2012.

To cite this article: Kerry Woods & Patrick Griffin (2013) Judgement-based performance measures of literacy for students with additional needs: seeing students through the eyes of experienced special education teachers, Assessment in Education: Principles, Policy & Practice, 20:3, 325-348, DOI: 10.1080/0969594X.2012.734777

To link to this article: http://dx.doi.org/10.1080/0969594X.2012.734777

Assessment in Education: Principles, Policy & Practice, 2013
Vol. 20, No. 3, 325–348, http://dx.doi.org/10.1080/0969594X.2012.734777

Judgement-based performance measures of literacy for students with additional needs: seeing students through the eyes of experienced special education teachers

Kerry Woods* and Patrick Griffin

Assessment Research Centre, Melbourne Graduate School of Education, University of Melbourne, Parkville, VIC, Australia

(*Received 21 May 2012; final version received 25 September 2012*)

This article describes the development of judgement-based performance measures to support the instruction of students with additional learning needs. The focus of the research was the design of assessment materials and protocols to help teachers recognise and respond to students' proficiency in foundational literacy skills. It drew on the expertise of special education teachers to provide all teachers with an evidence framework against which to observe their students' learning. The assessment materials were trialled in 53 schools and used to monitor literacy learning for 547 students, who ranged in age from 3 to 18 years and represented children and young people with a wide diversity and severity of disabilities. The article reports a new approach to judgement-based performance measurement which directs teachers' observations to meaningful shifts and transformations in foundational literacy skills for students with additional needs.

Keywords: teacher judgement; special education; literacy assessment; criterion-referenced performance measurement

Introduction

In 2002, an Australian federal government inquiry raised concerns about the social justice of provision of learning opportunities for students with intellectual and other forms of disability (Commonwealth of Australia 2002). In particular, evidence was cited of a deficit of knowledge about special education among teachers in mainstream schools and a pressing shortage of teachers with formal training in this field of education. Then, as now, the inclusive education policies endorsed by Australian state governments acknowledged the right of all students with disabilities to access schooling without experiencing discrimination (Australian Attorney General's Department 2005). The success of such policies, however, depended in no small part upon investment in the professional knowledge of teachers (Commonwealth of Australia 2002).

In the Australian state of Victoria, the location of the study described in this article, teachers in mainstream schools routinely made judgements about their students' learning progress against a standards-referenced curriculum framework – the Victorian Essential Learning Standards or VELS. These standards described the

*Corresponding author. Email: k.woods@unimelb.edu.au

skills and understanding considered essential foundations for further education, work and life, and the developmental levels through which most students are expected to advance (Victorian Curriculum and Assessment Authority 2006). In 2009, they were extended to include curriculum advice for school-aged children and young people with disabilities who were working at pre-foundational standards (Victorian Department of Education and Early Childhood Development [DEECD] 2009). However, that information was not widely used to monitor learning progress for students with additional needs. Victorian schools were required to develop an individualised learning plan for every student who received government-funded support on the grounds of disability (Victorian Auditor General 2007), but without clearly-mandated procedures to guide the process. Indeed, this idiographic strategy relied for its success on the knowledge of those charged with responsibility for devising a particular student's programme. In practice, its effectiveness was too often limited by teachers' lack of experience working with students with disabilities, and this was especially the case in integrated mainstream classrooms or smaller schools (Commonwealth of Australia 2002). Further, little or no advice was available to guide planning for students who experienced difficulties with learning, but who did not meet formal requirements for government-funded support. In contrast to the situation for non-disabled students, there was a lack of systematic information upon which teachers could draw when planning instruction for their students with additional needs (Commonwealth of Australia 2002).

Accordingly, the aim of the study described in this article was to provide teachers with a framework to support their understanding of the learning needs of their students with disabilities. More specifically, its purpose was the development of assessment materials to help teachers become more sensitive and competent observers of these students. It was designed to summarise teachers' judgements about students' proficiency into a reporting format that could be drawn on to devise, implement and monitor tailored programmes of instruction. The study focused on foundational skills related to reading, writing and other forms of symbolic representation. It was part of a larger programme of investigation into standards- or criterion-referenced assessment of learning for students with additional needs (Coles-Janess and Griffin 2009; Roberts and Griffin 2009; Woods 2010; Woods and Griffin 2010). As such, it built on the work of researchers who have developed criterion-referenced assessment frameworks to guide teachers' identification of students' burgeoning literacy skills (e.g. Annandale et al. 2003; Dewsbury 1994; Griffin, Smith, and Ridge 2001) by detailing evidence for the choice of assessment items and focusing on literacy learning for students with additional needs.

Traditionally, efforts to describe educational potential for students with additional needs, and to assign them to specialised programmes of instruction, have been based on age-based or norm-referenced interpretations of assessment. However, doubts have been expressed about the suitability of these forms of reporting for the purpose of planning educational programmes (e.g. Dacey, Nelson, and Stoeckel 1999; Jenkinson 1996; Vygotsky 1929/1993). For example, in his seminal work on the education of students with disabilities, Vygotsky questioned the usefulness of attempts to quantify how far these students fell behind their age peers. For Vygotsky, the limitations of such an approach were its negative emphasis upon impairment and loss, rather than abilities and compensation, and its conceptualisation of development as an additive, rather than a transformational, process. Instead, Vygotsky argued that educators should look for a student's strengths and capabilities, and then strive to draw on

those strengths to improve learning. Further, he proposed a way of thinking about learning, not as quantitative growth or the gradual strengthening of mental ability, but instead as a 'chain of metamorphoses' (42) or transitions from one qualitative type to another. This concept of learning as a series of qualitative transformations, which Vygotsky illustrated by describing transitions from crawling to walking, or from babble to speech, has since been fruitfully explored by many researchers whose work has been influential in the fields of education and developmental psychology (e.g. Anderson and Krathwohl 2001; Bruner 1983; Gagne 1985; Inhelder and Piaget 1958; Piaget 1947/2001). These researchers drew attention to the way that the behavioural expression of skill changed with increasing proficiency, although the overarching construct or skill domain remained the same. For example, Bruner described developmental transformations in a child's communication skills in terms of modes of representation, from the enactive (based on action within the immediate sensory environment), to the iconic, to more abstract, symbolic forms. Bruner's position was grounded on Piaget's seminal theory of the development of cognitive operations, progressing from the sensori-motor, to concrete operations, to the development of symbolic thought and formal operations. Thus, a complex skill domain such as literacy might be meaningfully conceptualised both as a continuum of ability and as a series of shifts or transformations through increasingly sophisticated levels of understanding and performance.

Assessment materials based on teacher observation and judgement

The materials developed in the current study were intended to provide a medium through which teachers' observation of the progress of students with additional needs could be coherently summarised, and then drawn upon as the basis of decisions about ways to support further learning. Principally, the assessment materials were intended to assist teachers to form decisions about students with a wide range of functional abilities, many of whom were expected to have disabilities of a severity or type that compromised their direct participation in testing procedures. For this reason, student response to assessment tasks or test items was precluded as a means of collecting information on their learning. Instead, the assessments were designed as a means of eliciting teachers' observations of their students in the unthreatening and authentic context of everyday classroom interactions.

Several researchers have commented upon the accuracy of teachers' observation-based predictions of performance on tests of reading and language achievement for their students with disabilities (e.g. Gresham, Reschly, and Carey 1987; Leinhardt 1983; Silverstein et al. 1983). In a similar vein, Griffin (1993) defended the use of teachers' professional judgement and direct classroom observation in student assessment. He noted that experienced teachers develop a rich store of knowledge about their students. They often internalise this knowledge, and their judgements of student performance become increasingly automatic. In practice, this means their interpretations of students' classroom behaviours may be under-valued as a source of information on student proficiency. Instead, Griffin argued that teacher observation should be accorded value as 'direct documentation' of student performance, and compared in favourable terms to the 'indirect estimation' (3) provided by students' scores on tests. However, Griffin's support for teacher judgement as a source of information on student proficiency builds on an expectation that teachers know their students well and also have a general schema or framework

against which to evaluate their performance. In the case of students with disabilities who are integrated into mainstream classrooms, the latter stipulation may too often be compromised by teachers' general lack of experience working with students at comparable stages of proficiency.

Further, the reliability and validity of assessments based on teacher judgement have been challenged due to concerns over lack of judgement objectivity (e.g. Bailey 1993; McCloskey 1990; Neisworth and Bagnato 1988) and poor consistency between observers (McCloskey 1990). These are undoubtedly areas of weakness for assessment procedures based on observer judgement, but weaknesses that can be ameliorated by training observers (Connally 2002; Griffin and Burrill 1995) and, in particular, training them in the use of a well-defined framework to guide assessment decisions (Gentile 1992). A fundamental aspect of the current study was thus an effort to develop materials based on such a framework, drawing on the knowledge and insights of experienced special education teachers and their capacity to accurately observe and interpret their students' behaviour.

Increasing the number of people making observations, and fostering moderation between observers, have also been linked to improved quality of judgements made in the context of special education and early intervention for students with learning difficulties (Bagnato et al. 2006). This draws notice to the importance of formal procedures that direct and organise teacher observation and, thereby, enhance both the reliability of judgement-based assessments and their efficacy at improving student learning. Moderation of teacher judgement and a team-based approach to assessment and decision-making were thus advocated in the current study in order to support and improve teachers' planning.

The decision to base assessment of proficiency upon teacher judgement and classroom observation led to identification of a number of constraints on the assessment materials. First, all items needed to be written in language that was unambiguous for teachers and that described behaviours teachers might reasonably be expected to detect in everyday classroom interaction with students. Second, it was important that teachers have sufficient opportunity to observe students before using the materials. Third, the assessment materials needed to be concise and easy for teachers to use, while also providing sufficiently detailed information about students to support understanding of their current level of proficiency, and to guide decisions about how best to structure the learning environment and tailor instruction.

Building a framework to support teacher judgement of student performance

The method adopted for the design of the assessment materials was based on a procedure for defining standards- or criterion-referenced frameworks (Griffin 1993, 2007), which relies on the collaboration of subject-matter experts, working within the format of a partial credit latent trait model (Masters 1982), to define the relative discriminating power of components of complex observation structures. Such an approach to educational measurement assumes that a construct of interest can be described as a continuum with direction and units of magnitude, and meaningfully defined by cohesive sets of behaviours that represent levels of increasing proficiency (Griffin 2007) or, indeed, transitions from one qualitative expression of proficiency to another (Vygotsky 1929/1993).

With this in mind, the study invited experienced teachers and consultants in the fields of special education and literacy learning to propose, critique and expand

upon a hypothesised developmental framework for foundational literacy skills. The first stage in this process was the development of a jointly agreed definition of the skill domain. The intention was to specify the construct in a manner that could be used to develop rubrics or scoring criteria (e.g. Gronlund 1998; McDaniel 1994) through the description of a set of representative capabilities that could, in turn, be expressed in terms of readily observable student behaviours.

Rubrics were then developed by a group of subject-matter experts drawn from a cross-section of academic backgrounds, representing a variety of perspectives. They were defined as, 'a set of scoring guidelines that describe the characteristics of the different levels of performance used in scoring or judging a performance' (Gronlund 1998, 225). The central features of rubrics are the ordered categories or levels of performance that comprise a description of the cognitive, affective and psychomotor skills embedded in competent performance (Griffin 2007). Underpinning the concept of rubrics is the criterion-referenced interpretation in which an individual's achievement or competence is described in terms of the behaviour they demonstrate or the tasks they can perform (Glaser 1981). The use of criterion-referenced definitions for rating scales conveys greater information about the quality of performance, discriminates more accurately between individuals and allows for students to be given diagnostic feedback that they will likely perceive as constructive and valid (Bondy 1983).

The rubrics were to be presented in a questionnaire format, in which items consisted of a performance indicator as the stem and quality criteria as ordered alternative choices. The order of increasing quality performance in the quality criteria is important in that a partial credit model (Masters 1982) can be employed to integrate the responses from the multiple observers over multiple items. In that model, the probability that person n is scored as or responds in category x to item i is given by:

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x_i} (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij})}$$

where x takes the values 0,1,2,3 ... m_i corresponding to 'm_{i+1}' ordered categories associated with item i; 'β_n' is the ability of the person n, and δ_{i1}, δ_{i2}, δ_{i3}, δ_{i4}, ..., δ_{im}, are the m_i, difficulty parameters associated with the thresholds for item i or the thresholds on continuum i. The Rasch model then simultaneously allows person ability and item difficulty to covary, and the number of thresholds (m_i) to vary from item to item or from continuum to continuum. That is, the number of ordered categories can vary across items. This is a powerful freedom from constraint on the number of categories that is often perceived to be operating in the construction of rubrics.

Construct definition – representative capabilities and behavioural indicators

To define the construct of foundational literacy, it was first acknowledged to comprise a complex set of capabilities that change in their expression with improved proficiency (Chall 1967, 1983; Clay 1967, 1991; Spear-Swerling and Sternberg 1996). These include, for example, language skills, motivation to participate in literacy activities, awareness of symbols, knowledge of the conventions of

print, and alphabetic, orthographic and phonological awareness (e.g. Clay 1991; Ehri 1992; Frith 1985; Holdaway 1979; Spear-Swerling and Sternberg 1996). In addition, there are particular capabilities, such as manual control and dexterity, which underpin development of a functional writing system (McCutchen and Berninger 1999). Yet, among students with additional needs, impaired mobility may require that assistive technologies be used to enhance, or replace, skills requiring dexterity. It was, therefore, acknowledged that a comprehensive measure of proficiency would need to recognise these alternate modes of expression. It was also deemed important to note that some students may not learn to read or write in the conventional sense, but instead develop alternative methods for sending, receiving and organising information through the use of non-text symbols and pictures, or a combination of these forms and conventional text. An important consideration was thus the inclusiveness of the assessment materials, and their fairness and relevance for students with a wide diversity of modes of expression.

Accordingly, the construct of interest (i.e. foundational literacy for students with additional needs) was initially defined as *the process of building and conveying meaning through written or produced symbols and text.* This definition was chosen because it did not privilege the use of a particular medium or sensory system. It could encompass diversity in the means by which proficiency is demonstrated, while retaining emphasis on literacy as a way of conveying and comprehending ideas and information presented in symbolic form. It was described in terms of a representative set of capabilities, which were then drawn upon by subject-matter experts to draft a set of indicative and observable behaviours for each capability (Woods 2010). These are listed in Table 1.

Thus, each of the abstract capabilities selected to represent the construct was defined in terms of a set of observable skills or indicative behaviours that subject-matter experts identified as appropriate for that capability. The next step in this process, working towards development of a pool of assessment items, was the specification of criteria of performance quality for each of the indicative behaviours. The steps taken in this process are set out in Figure 1, with more detailed explanation provided in the section that follows. For readability, this diagram describes the construct in terms of only four capabilities when, as shown in Table 1, six broad areas of interest were identified.

Identification of quality criteria

As Vygotsky (1929/1993) argued, developmental measurement should ideally describe transformations in performance quality that mark increasing proficiency for students. In other words, measurement of progress should not only acknowledge *what* a student needs to learn (in terms of critical capabilities and the observable behaviours that stand as demonstration of those capabilities), but also describe *how well* each of those behaviours is performed as proficiency improves. Crucially, information derived from an assessment should be interpretable for teachers as a marker that a student is ready for a new challenge or an expanded set of learning experiences. This information should be able to guide teachers as they strive to target instruction for students, so that learning activities are pitched at levels that are neither too high nor too low for the student.

With this in mind, experienced special education teachers were asked to describe three or four possible descriptors of performance quality for each of the behaviours

Table 1. Assessment items: capabilities and behavioural indicators.

Literacy – Building and conveying meaning through written/produced symbols	
Representative capabilities	Draft behavioural indicators
Awareness of symbols and print	• Responding to symbols • Matching symbols and meaning • Ordering symbols to express ideas • Responding to photographs and pictures • Knowing how to handle reading materials • Recognising that print has a consistent meaning • Using the terminology of printed text • Using punctuation in writing • Using basic grammar in writing
Motivation to participate in literacy activities	• Participating in reading • Enjoying and relating to reading • Choosing books and stories • Showing interest in drawing or writing • Selecting materials for drawing or writing • Showing pride in drawing or writing
Knowledge of letters and sound–letter relationships	• Identifying letters and numbers • Matching letters to sounds • Matching words on the basis of alliteration • Matching words on the basis of rhyme • Blending sounds in words • Segmenting sounds in words
Orthographic knowledge	• Recognising word forms • Recognising parts of words
Comprehension	• Predicting the meaning of words • Predicting the content of reading material • Re-telling stories
Control of production	• Producing letter forms • Copying words • Knowing how to present written material • Using pens/pencils for writing • Using computer keyboard and mouse for writing

listed in Table 1. These were written to capture the range of proficiency that teachers expected to observe among their students. For each of the indicative behaviours, teachers were asked to bring to mind students with different levels of competence, and to describe the sorts of behaviours they would expect these students to display. This is illustrated in Table 2, which presents a set of performance quality criteria for one of the items. Teachers were asked to imagine a student who demonstrated the lowest observable level of a particular behaviour, then to think about a student

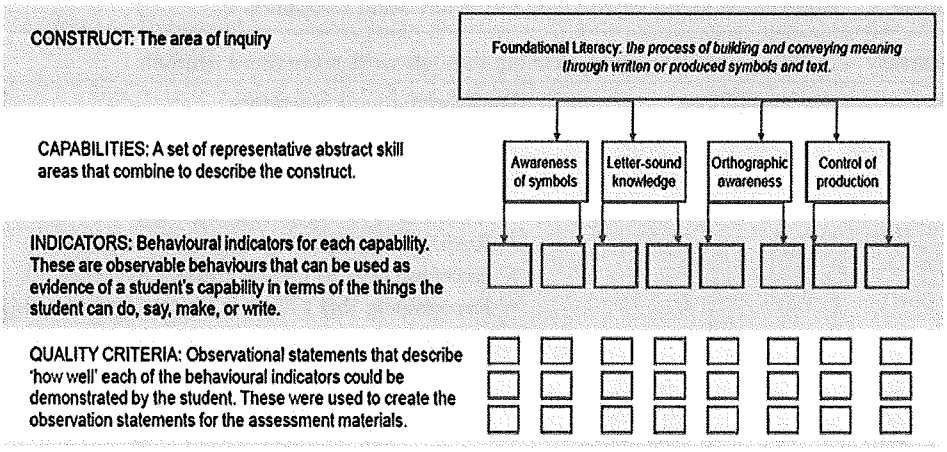

Figure 1. Stages in the development of an item pool for the assessment materials.

Table 2. Example of performance quality criteria.

Capability: awareness of symbols and print	
Indicative behaviour	*Quality criteria*
Responding to symbols	(1) Looks at symbols and/or points to them (2) Matches a symbol to an object, animal, person or activity (3) Identifies and names symbols and words (4) Interprets ordered sequences of words, symbols or signs to understand messages

who had achieved mastery of the behaviour and then to describe the major transitions in learning they expected to observe between the lowest and highest levels. They were challenged to articulate the specific behaviours they would watch for and use as markers that a student was ready to move ahead in their learning and should thus be provided with a new set of learning experiences.

Development of a hypothesised standards-referenced framework

There are many recommended procedures for setting performance standards (e.g. Berk 1996; Cizek 2006; Jaeger 1995). Typically, it is an activity conducted relatively late in the process of instrument development, when statistical information about student performance on assessment items is available for scrutiny (Wolfe and Smith 2007). However, Griffin (2000) proposed the use of a judgement-based application of the partial credit model (Masters 1982) to develop a hypothesised standards-referenced framework for an instrument. This exercise draws on the knowledge and experience of subject-matter experts to describe *a priori* expectations of item difficulty ordering. It can be used to indicate how well draft items or, as in the current study, the performance criteria drafted for items, collectively reflect the construct of interest, and to draw attention to possible gaps, redundancies and

inconsistencies in assessment materials. Further, hypothesised proficiency standards derived from such an exercise can be compared to those observed from later statistical calibration of the instrument, and thus contribute to arguments for or against its validity. This comparison can be used to evaluate how well the assessment instrument fits its intended purpose.

The approach followed procedures used to identify profiles of literacy learning for non-disabled students (Griffin et al. 2001), and involved the conduct of a content analysis in a manner similar to the interpretation of factor meaning in a factor analysis. It used a method in which the draft quality criteria for each item were ordered and placed on a matrix by subject-matter experts. In the current study, this group of experts included seven teachers from special education schools, each with more than 10 years of experience in this field of education, two literacy curriculum coordinators from mainstream primary schools and three academic researchers in the fields of literacy learning and special education. Over two days, they worked in pairs to examine each of the draft items and its quality criteria in terms of the expected proficiency required for their performance, and to place them on a matrix with the criterion judged as easiest located at the bottom of the matrix and criteria judged as progressively more difficult positioned at correspondingly higher levels. Pairs worked independently, then compared and discussed their judgements as a collective until a consensus position had been reached. These steps were repeated for each item, comparing its criteria with the positioning of those of the first, and then subsequent, item(s) to determine their relative placement in terms of expected difficulty. The result was a matrix of 32 draft items expressed as behavioural indicators, each with between two and four criteria of performance quality. To illustrate the process, a section of the matrix showing seven example draft items is included as Appendix 1.

As the draft quality criteria were positioned on the matrix, subject-matter experts worked independently, and then as a group, to examine clusters of criteria located at the same, or similar, levels and over time to reach agreement on whether meaningful interpretations of those levels could be discerned. This led to a hypothetical definition of a learning continuum in terms of standards of increasing proficiency, as shown in Table 3.

The outcome of this work was the development of a pool of draft assessment items, which were then piloted with 13 experienced special education teachers and four literacy or disability coordinators in one mainstream primary school and one mainstream secondary school. At the completion of the piloting, slight revisions were made to the wording of some draft items and their quality criteria to improve clarity of expression and interpretability for teachers. This led to a final selection of 30 items to be tested in schools.

A large-scale field test of items was then conducted in 78 schools (56 specialist schools and 22 mainstream schools). Six-hundred and seventy-four teachers, or in many cases teams of teachers, responded to draft items presented to them on scannable paper forms. In the main, teachers described proficiency for between one and four of their students, such that the draft pool of items was tested with a group of 1646 students aged between 3 and 18 years and with a diversity and range of additional learning needs (Woods 2010). An example of one of the items is shown in Figure 2.

Each item was thus written as an observation statement with a set of possible response options corresponding to the quality criteria drafted by subject-matter

Table 3. Hypothesised literacy levels in a standards-referenced framework.

Description of standard	Examples of representative quality criteria
Application and extension of literacy knowledge	Decodes unfamiliar words using known clusters of letters and spelling patterns. Adapts presentation of written material to suit a range of purposes and audiences
Recognition and use of conventional spelling patterns and rules of text presentation	Predicts meaning of words using all or most letter information. Matches sounds to clusters of letters and spelling patterns that are common in English. Knows how to use upper and lower case forms of letters in written work
Manipulation of sounds in words and rudimentary recognition of conventions of text presentation	Produces sounds to match common letter blends. Sorts and matches objects that have the same first sound. Copies words from modelled examples. Writes or types from left to right across page and from top of page to bottom
Awareness and use of relationships between letters and sounds	Names all letters and identifies most common sounds. Uses the sounds of letters/symbols in invented spelling and writing. Predicts meaning of familiar words using illustrations and partial cues
Use of visual cues to recognise and label symbols, including some letters and familiar words	Identifies and names photographs and pictures. Names some letters. Sorts and matches letters and numbers. Recognises some very familiar words by sight. Draws, scribbles or writes letter forms, mixed with numbers and shapes
Pre-alphabetic, responsive to pictures, shapes and sounds	Looks and points at realistic photographs of objects. Sorts and matches pictures and shapes. Imitates the sounds of letters when prompted. Draws non-linear shapes and forms
Responsive to objects within immediate sensory environment	Uses books for sensory/exploratory activities. Remains present while a story is being read. Makes choices between objects (or photographs of objects)

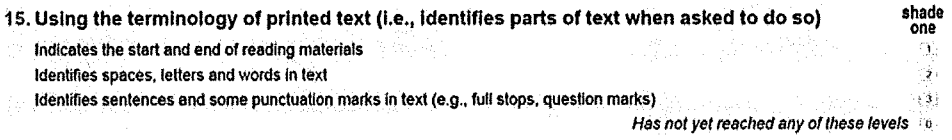

Figure 2. Example of one draft item from the pool tested in schools.

experts. Each criterion within a set was ordered and numbered, with a code of 1 denoting the lowest observable level of performance and a code of 0 to be used to indicate that a student had not yet demonstrated that behaviour. The number assigned to each criterion was an ordered code, indicating the sequential difficulty within an item. There was no assumption made about the equivalence or non-equivalence of criteria with the same numerical code on different items, nor of the magnitude of difference between quality criteria within an item.

Following data collection in schools, items in the draft pool were calibrated using a (Rasch 1960/1980) model for partial credit scoring (Masters 1982). This permitted an initial observation that the spread of item difficulties was a good match

to the range of abilities of a large sample of students aged from 3 to 18 years, and drawn from both special education and mainstream settings (Woods 2010). In general, overall fit to the model was good, and person and item separation reliability indices of 0.98 and 0.99, respectively, supported a judgement that the draft items were able to separate students in terms of their ability (Wright and Stone 1999), and were spread out well along a continuum of proficiency. Some individual items were flagged for revision on evidence of misfit to the Rasch model or non-sequential ordering of score categories (Woods 2010).

In addition, there was considerable redundancy in the substantive information available from the scaled locations of quality criteria, which indicated that it was possible to split the item pool to develop alternate and linked forms of the assessments (Woods 2010). The development of alternate forms of an assessment is commonly motivated by concerns about test security or, as in the current study, situations in which assessments are intended for administration at different points in time to allow teachers to track and reflect upon student learning progress (McDonald 1999). The intention was to develop two shorter forms of each set of assessment materials, each comprising 20 items of which 12 were shared, and which could be linked using an item response modelling technique for common item equating.

The next phase of the research involved a study of the use of the revised assessment materials in schools, and statistical calibration of the assessment items to build an empirically-derived progression of literacy learning for students.

Use of the assessment materials in schools

Schools and students

Teachers of 547 students in 53 Victorian government schools (43 specialist schools and 10 mainstream schools) participated in a trial of the revised form of the assessment materials. The students ranged in age from 3 to 18 years, and approximately 64% were boys. All of the students had a form of intellectual or developmental disability, and 96% were identified by their teachers as having specific additional needs in terms of their language and literacy skills. Twenty-seven per cent of the students had a diagnosis of autism spectrum disorder and 25% had restricted or impaired mobility. The participant group of students also included a small number who had particular co-occurring conditions such as cerebral palsy (3%), or severe health impairment (4%), or who were blind (3%), or deaf or hard of hearing (2%).

Further, 38% of the students were enrolled in special developmental schools for students with moderate to profound intellectual disabilities, 26% attended schools for students with moderate to mild intellectual disabilities, 16% attended rural schools that enrolled students with a broad range of disabilities, 5% attended schools for students with autism spectrum disorder, 12% attended mainstream primary schools and 3% attended mainstream secondary schools.

Teachers in each of the schools worked within the supportive framework of a collegial group to use the assessment materials and review their students' progress from the middle to the end of one school year. Working in small teams, which could include their teaching colleagues and support professionals such as speech therapists or disability coordinators, they responded to each item on the assessment materials by selecting the criterion that, in their judgement, was the best match to a particular student's everyday classroom behaviour. The number of people who used

the assessment materials varied in each school, and was not recorded in this phase of the research.

Calibration of the assessment materials

Responses to items on the assessment materials were coded as rating-scale ordered categories, and calibrated using a Rasch (1960/1980) model for partial credit scoring (Masters 1982). A map of the variable, taken from analysis using the ConQuest item response modelling software (Wu, Adams, and Haldane 2008), is shown in Figure 3. This displays the distribution of students (represented on the map by Xs) and item quality criteria (represented by numerical codes that show the item number and quality criteria or step score, such that 1.1 indicates the first criterion on the first item, 1.2 the second criterion on the first item and so on). Students and item quality criteria are mapped on the figure against a logit scale, which can be interpreted as an indication of the demand or difficulty of each criterion.

Alternate and linked forms of the assessment were used to monitor learning for a matched sample of students across six months, and equated using the ConQuest

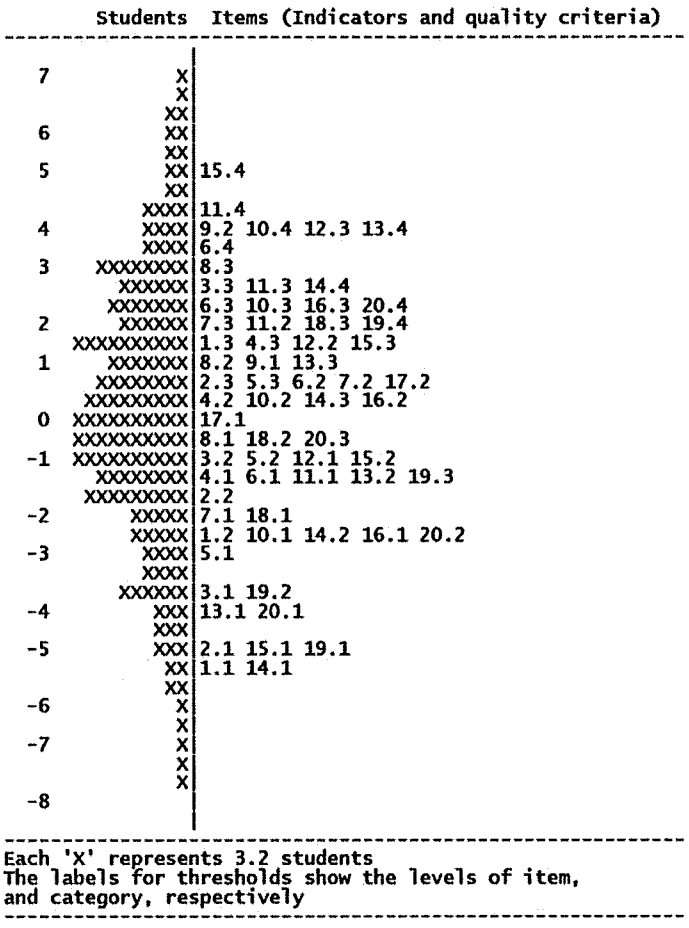

Figure 3. Variable map for the literacy assessment items (indicators and their performance criteria).

item response modelling software (Wu et al. 2008), and a Rasch method for equating tests with common item anchoring. Equating is a procedure used to bring different measures of the same construct into alignment, such that all items in a pool can be described with the same unit of measurement. It is a method used to ensure that student performance on different subtests of items can be mapped onto a single continuum and meaningfully compared.

Each of the alternate forms had an alpha reliability of 0.98, indicating that each set of items displayed strong internal consistency, and item and person separation reliabilities of 0.99 and 0.97, respectively. Information-weighted mean-square residual goodness of fit (infit) statistics ranged from 0.81 to 1.47 for the revised items, and difficulty thresholds (δ) were ordered for all items. The value of infit mean-square statistics is 1.0 when the data are a perfect fit to the underlying Rasch model. Wilson (2005) pointed out that there is no absolute limit to judge a good weighted mean-square value, and noted that previous researchers suggested a lower bound of 0.75 and an upper bound of 1.33 as very approximate guides to judgements of model fit (e.g. Adams and Khoo 1996). Wright et al. (1994) argued that reasonable ranges for fit statistics depended upon the type and purpose of measurement, and recommended a range of 0.5 to 1.7 for clinical observation compared with 0.8 to 1.2 for a high-stakes, multiple-choice test. They proposed that items with fit statistics with values >2.0 distort or degrade measurement, items with values between 1.5 and 2.0 may be unproductive for measurement construction but are not degrading, items with values between 0.5 and 1.5 are productive for measurement, and items with values <0.5 are less productive for measurement, but not degrading, although they may produce misleadingly high indices of reliability.

Next, maps of the variable (e.g. as shown in Figure 3) were used to determine initial clusters of item quality criteria. Each cluster was then reviewed by a group of experienced special education teachers. This was a different group of teachers to those who had assisted with development of the assessment materials. Their challenges were to confirm or dispute the clustering of quality criteria, and then to identify and summarise the substantive meaning of behaviours that clustered at successive levels of estimated difficulty, and to reflect on whether these could be meaningfully related to programmes of teaching intervention for students at different levels of proficiency. An example of one of the clusters is shown in Table 4, with the letters A and B used to show that a particular quality criterion appeared on the first (A) and/or second (B) administration of the assessment instrument.

Seven such proficiency levels were identified. For example, quality criteria clustering at the lowest level of proficiency described students who were beginning to engage with objects and photographs of familiar objects or people and, at the next level, they described students who were learning to identify pictures, shapes and sounds, and beginning to role-play reading and writing activities. Quality criteria at the third level described students who were becoming aware of print in the environment, and learning to use visual cues to recognise letters of the alphabet and some very familiar words. At the fourth level, they described students who were beginning to understand that letters relate to sounds in systematic ways and who were learning to use symbols as tools for self-expression and communication. Quality criteria at the fifth level described students who were learning to use relationships between letters and sounds in their attempts at spelling and writing. Those that clustered at the sixth level described students who were learning to recognise and use

Table 4. Quality criteria that clustered at the lowest level of literacy proficiency.

Criterion no.	δ	Description
1.1AB	−5.24	Looks at, touches or pats photographs of familiar objects
14.1AB	−5.09	Responds to photographs of familiar objects/people (e.g. smiling, touching)
15.1AB	−5.04	Makes choices between objects (or photographs of objects)
2.1A	−4.88	Remains present while a story or other reading material is being read or shown
19.1AB	−4.74	Picks up and holds objects
20.1AB	−4.03	Taps an object with a finger
13.1A	−4.01	Accepts materials for drawing or writing
3.1AB	−3.76	Shows enjoyment of being read to (e.g. by smiling, looking, relaxing)
19.2AB	−3.68	Holds and uses large crayons and pencils, perhaps with a fist-like or similar grip

Notes: Criterion numbers are used to denote both the order of items on the assessment materials and the order of the quality criterion within an item, such that criterion 1.1 refers to the first quality criterion for the first item. The letters AB indicate that an item with its quality criteria was included on both parallel forms of the assessment materials. Where a quality criterion is identified with only the letter A or B, it appeared on only one of the parallel forms.

patterns of spelling that are common in English, and to respond to information derived from reading. At the highest level, quality criteria described students who were learning to use a range of strategies to confirm or modify their understanding of written material, and to apply this understanding in a wide range of literacy activities.

The empirically derived progression was then checked against the original, hypothesised framework. There was a very close match between the expected difficulty order mapped during the development of the item pool and the observed difficulty of item quality criteria. Indeed, the Spearman correlation between observed levels of difficulty and subject-matter experts' expectations of proficiency levels required for each quality criterion showed a strong relationship ($\rho = 0.95$). This was interpreted as evidence for the construct or substantive validity of the instrument (Wright and Stone 1999).

A study of teacher judgement and planning

Statistical analyses can be used to reflect upon the technical quality of data derived from the use of an assessment instrument. However, a comprehensive consideration of the validity of assessment materials should also encompass consequential aspects (Messick 1995) that, for example, use assessment outcomes as a starting point for teachers' planning and implementation of learning programmes. This argument for validity proposes, quite simply, that an assessment is valid if it not only measures the construct it was intended to measure, but also supports the sorts of decisions it was intended to support. A further round of qualitative research was thus conducted in 12 schools to observe and document teachers' capacity to use the assessment materials, with emphasis placed on interpretation of student results by teachers and

their efficacy at informing decisions about how best to target and tailor learning experiences.

The schools taking part in this final stage of the research included five for students with moderate to severe intellectual disabilities, three schools for students with mild intellectual disabilities, two schools for students with autism spectrum disorder and two mainstream primary schools. In each school, teachers organised themselves into teams of four to five colleagues to review student assessment data and maintain logs of their plans and decisions. At the outset of the study, representatives from each school were guided through protocols for the use of developmental assessment, as described by Griffin et al. (2010). These protocols were developed from research into evidence-based teaching and assessment processes conducted within a framework of developmental learning, and the expectation that such processes could enhance learning outcomes for students (Griffin et al. 2010). Teachers were thus provided with advice on how to use and interpret their students' assessment data. Protocols were established for teachers to make and record decisions, debate and review the impact of teaching strategies and monitor learning for their students.

Teachers tracked learning progress in foundational literacy skills for their students within one school year. They observed their students' typical behaviours in everyday classroom interactions and then responded to an online version of the literacy assessment materials. Teachers could download student reports as soon as they completed an assessment. An example of an individual student report is shown in Figure 4. The format for this report was derived from Griffin's (2001) work on reporting practices that foster teachers' capacity to set realistic goals and targets for student learning.

Reports were thus presented in the form of a developmental progression, with proficiency levels or standards displayed in hierarchical order. Each successive level was described by a generalised description of its meaning shown adjacent to the level. These descriptions were summaries of the proficiency standards derived from calibration of the assessment materials and the interpretations of clusters of quality criteria made by experienced special education teachers.

Records of teachers' collaborative decisions about teaching programmes for their students were returned to the researchers for analysis. Interpretation of these records revealed some common themes in the way experienced teachers typically planned learning experiences and support for their students. The first of these placed importance on the need to provide a multidisciplinary approach to programme delivery, in which classroom teachers worked closely with therapy staff, student welfare staff, support staff and their teaching colleagues. The second theme emphasised the importance of building and maintaining positive and caring relationships with students and their families, and drawing on these relationships to personalise students' programmes. The third built on this idea to stress the fundamental importance of knowing each student well, understanding their likes, dislikes and preferences, and the materials or topics that were likely to interest and engage them.

Analysis of teachers' records also showed that experienced special education teachers rarely referred to the specific nature of a student's disability when setting targets for teaching and learning. Indeed, the nature of a child's disability and other personal characteristics, such as their age or gender, were rarely mentioned in planning documents, except in terms of the sorts of resources teachers drew upon when working with the student. For example, many teachers reported that they tailored

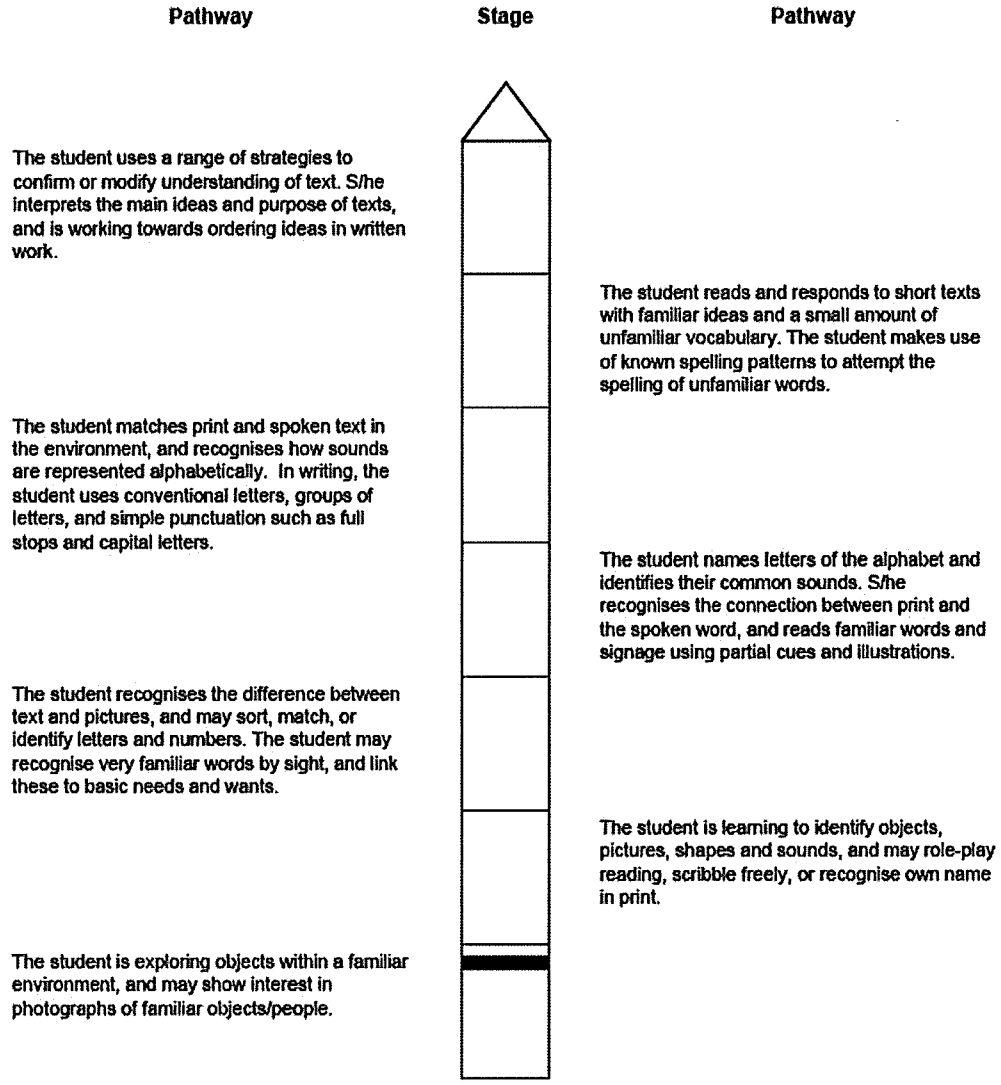

Figure 4. Individual report of student proficiency in foundational literacy skills.
Note: The student is estimated to have reached the level represented by the black line.

and personalised the materials and books they used to closely match the student's interests, and to include representations of the student and favourite or familiar items, food, people and activities. This was particularly the case for students at the lower levels of proficiency.

Other teaching strategies and resources widely used for students at lower proficiency levels included multi-sensory activities, music and songs, games and puzzles, information and communications technology (ICT) resources such as interactive whiteboards and interactive computer programmes, choice-making activities, augmentative communication systems and pictorial-aided language displays, a print-rich environment, daily individual and small group reading sessions and a range of pre-writing activities to build strength and coordination. Most teachers also recorded the

importance they placed on the establishment of structured and predictable classroom schedules and routines for their students, in many cases presented in the form of visual timetables. Most used a form of daily communication diary to share information with their students' family members.

Teachers adapted the materials used to challenge students as they improved their general level of proficiency. Most teachers noted the importance of literacy resources that were appropriate for students at different age levels, and the challenge of sourcing materials that were well-matched to the student's proficiency level in reading but not targeted towards very young children. Teachers relied upon the internet and printing and laminating resources to personalise literacy materials for their students, and, for students at all levels of proficiency, resources such as interactive whiteboards, tablet personal computers or other ICT tools were strongly endorsed.

Teachers recorded their use of guided reading and writing activities, modelling new skills, and shared drafting and editing for students at higher levels of literacy proficiency. The use of literacy skills in authentic, everyday and practical situations was strongly emphasised, so that many teachers planned programmes of literacy instruction around excursions, cooking activities, shopping trips, sporting and recreation events, following directions and instructions, internet research, emailing and blogging to emphasise reading and writing as skills that had personal relevance and usefulness for students. Instructional programmes included a strong component of explicit teaching of the skills identified as targets for student learning, with the use of repetition to embed new skills, and then variation of context, purpose and amount of support provided to the student to gradually increase independence and capacity to generalise skills across different situations and tasks. More proficient students were introduced to thinking and planning tools that combined written and visual presentation of information, such as graphic organisers, task cards, timetables and personal schedules, and encouraged to make and use personal word banks or dictionaries.

As part of this investigation, student learning was monitored over one school year, with six months between assessments. Figure 5 shows the distribution of students at each level of literacy proficiency measured at two points in time, and indicates a general improvement of proficiency for students in the 12 participating schools in this time period.

The majority of students made sound progress in their literacy learning, with an average change of 0.51 logits ($SE = 0.07$ logits) over six months. As shown in

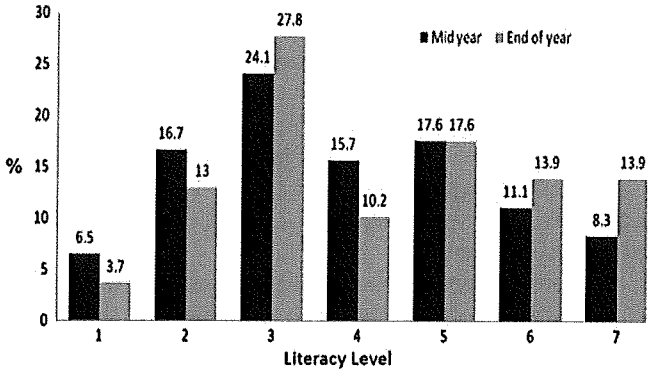

Figure 5. Progress in literacy proficiency (% students).

Figure 5, a considerably smaller number of students were working at lower proficiency levels at the end of the year in comparison with the mid-year assessment. Similarly, many of the students had progressed to higher overall levels of skill and understanding in that time.

Teachers' observations and comments

At the conclusion of the study, a teacher from each team in each participating school was interviewed about perceptions of the utility of the assessment materials and protocols. Across the 12 schools, this meant that 15 teachers were interviewed in their role and capacity as representative of teams each of four or five colleagues. Each teacher in each team had used the assessment materials to monitor learning progress for two or three students, within a context of collaborative, team-based moderation and decision-making (Griffin et al. 2010).

The consensus opinion among these representative teachers was that access to the assessment materials had improved their capacity to organise their thinking about their students' learning, plan in a more purposeful way, and target learning experiences for students and meet their specific learning needs with more sensitively tailored resources and activities. For example, one teacher commented that her colleagues often remarked that they 'did not know where to take the students next' and that they sometimes felt under pressure to set goals for student learning that were quite unrealistic or too general to inform programmes of teaching. According to the reports of representative teachers from each of the participating schools, the assessment materials were widely welcomed by others in their collegial teams as a means of directing attention to literacy behaviours that are important for students to establish, and summarising their observations into a reporting format that helped teachers make and share decisions about appropriate objectives for student learning. These materials have since been endorsed as part of mandated assessment, reporting and planning procedures for all students with intellectual and developmental disabilities in Victorian government schools (Victorian Department of Education and Early Childhood Development [DEECD] 2011). They have been released to schools as part of the Abilities Based Learning and Education Support (ABLES) materials for students with additional needs.

Conclusion

The aim of the research described in this article was the design and validation of a judgement-based performance measure to help teachers recognise proficiency in foundational literacy skills for students with additional learning needs. The study drew on the expertise of special education teachers, and their capacity to observe and document learning for their students, to provide all teachers with a framework against which they could monitor learning for their students. In short, the intention was to help teachers see their students with disabilities through the eyes of experienced special education teachers.

There were two assumptions about learning among students with additional needs that were foundational to the research. The first of these was a belief that every child is capable of learning, an assumption that drew on Vygotsky's (1929/1993, 80) contention that:

> [t]he greatest mistake – the view of a child's abnormality as only an illness – has made our theory and practice subject to a most dangerous delusion. No matter what the affliction may be. ... we meticulously analyse every corpuscle of the defect, every little speck of disease found in abnormal children, while we never notice the gold mines of health inherent in each child's organism.

The second assumption was that teachers could use student assessment outcomes presented in terms of developmental progressions to plan and implement programmes of learning for their students. In Vygotsky's terms, the primary purpose of the study was to draw notice away from the 'specks of disease' and, instead, concentrate the attention of educators on the 'gold mines of health' and potential in each student with additional needs.

Notes on contributors

Kerry Woods is a research fellow at the Assessment Research Centre in the University of Melbourne's Graduate School of Education. Her research interests centre on the design of assessment instruments and protocols to support teachers' instructional planning and the implementation of teaching programmes that are targeted and differentiated to meet students' learning needs. Her work has led to the development of an integrated programme of advice and support for teachers of students with additional learning needs.

Patrick Griffin holds the chair of education (Assessment) at the University of Melbourne, is director of the Assessment Research Centre and is the associate dean in the Melbourne Graduate School of Education. For more than 20 years, his work has focused on the application of item response modelling to criterion-referenced performance measurement. He has led many large-scale research and evaluation projects, both nationally and internationally, including the development of professional standards for classroom teachers and educational managers in Australia, Vietnam and China, and provision of evidence-based policy advice to support the education of students with additional needs.

References

Adams, R., and S.T. Khoo. 1996. *Quest*. Melbourne: Australian Council for Educational Research.

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

Annandale, K., R. Bindon, K. Handley, A. Johnston, L. Lockett, and P. Lynch. 2003. *First steps second edition: Linking assessment, teaching and learning*. Port Melbourne, VIC: Rigby Heinemann.

Australian Attorney General's Department. 2005. Disability Standards for Education. http://www.ag.gov.au/www/agd/agd.nsf/page/humanrightsandanti-discrimination_disabilitystandardsforeducation (accessed March 3, 2011).

Bagnato, S., J. Smith-Jones, M. Matesa, and E. McKeting-Esterle. 2006. Research foundations for using clinical judgement (informed opinion) for early intervention eligibility determination. *Cornerstones* 2, no. 3: 1–4.

Bailey, M. 1993. Judgement, evidence and the assessment of competency. Paper presented at Testing times: A national conference on assessment for competency-based training, in Adelaide.

Berk, R.A. 1996. Standard setting: The next generation (where few psychometricians have gone before!). *Applied Measurement in Education* 9, no. 3: 217–35.

Bondy, K.N. 1983. Criterion-referenced definitions for rating scales in clinical evaluation. *Journal of Nursing Education* 22, no. 9: 376–82.

Bruner, J. 1983. *Child's talk: Learning to use language*. New York, NY: W.W. Norton and Company.

Chall, J. 1967. *Learning to read: The great debate*. New York, NY: McGraw-Hill.

Chall, J. 1983. *Stages of reading development.* New York, NY: McGraw-Hill.
Cizek, G.J. 2006. Standard setting. In *Handbook of test development*, ed. S.M. Downing and T.M. Haladyna, 225–58. Mahwah, NJ: Lawrence Erlbaum.
Clay, M. 1967. The reading behaviour of five year old children: A research report. *New Zealand Journal of Educational Studies* 2, no. 1: 11–31.
Clay, M. 1991. *Becoming literate: The construction of inner control.* Portsmouth, NH: Heinemann.
Coles-Janess, B., and P. Griffin. 2009. Mapping transitions in interpersonal learning for students with additional needs. *Australasian Journal of Special Education* 33, no. 2: 141–50.
Commonwealth of Australia. 2002. *Education of students with disabilities, being a report of the Senate Education, Employment and Workplace Relations Committee.* Canberra: Australian Government Publishing Service. http://www.aph.gov.au/senate/committee/eet_ctte/completed_inquiries/2002–04/ed_students_withdisabilities/report/index.htm (accessed March 3, 2011).
Connally, J. 2002. Inter-rater agreement, rater leniency, and factors influencing judgements of higher order competencies with a multi-rater competency based assessment framework. Paper presented at the annual conference of the Australian Association for Research in Education, December 1–5, in Brisbane.
Dewsbury, A., ed. 1994. *First steps: Reading developmental continuum.* Education Department of Western Australia. Portsmouth, NH: Heinemann.
Dacey, C.M., W.M. Nelson, and J. Stoeckel. 1999. Reliability, criterion-related validity and qualitative comments on the fourth edition of the Stanford-Binet intelligence scale with a young adult population with intellectual disability. *Journal of Intellectual Disability Research* 43, no. 3: 179–84.
Ehri, L.C. 1992. Review and commentary: Stages of spelling development. In *Development of orthographic knowledge and the foundations of literacy: A memorial festschrift for Edmund H. Henderson*, ed. S. Templeton and D.R. Bear, 307–32. Hillsdale, NJ: Lawrence Erlbaum.
Frith, U. 1985. Beneath the surface of developmental dyslexia. In *Surface dyslexia*, ed. K.E. Patterson, J.C. Marshall, and M. Coltheart, 301–30. Hillsdale, NJ: Lawrence Erlbaum.
Gagne, R. 1985. *The conditions of learning.* 4th ed. New York, NY: Holt, Rinehart & Winston.
Gentile, C. 1992. *Exploring new methods for collecting students' school-based writing: NAEP's 1990 portfolio study.* Washington, DC: National Centre for Education Statistics.
Glaser, R. 1981. The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist* 36, no. 9: 923–36.
Gresham, F.M., D.J. Reschly, and M.P. Carey. 1987. Teachers as 'tests': Classification accuracy and concurrent validation in the identification of learning disabled children. *School Psychology Review* 16, no. 4: 543–53.
Griffin, P. 1993. *Profiles: Assumptions and procedures in their development.* Melbourne: Assessment Research Centre, RMIT.
Griffin, P. 2000. Competency based assessment of higher order competencies. Paper presented at the NSW ACEA state conference, April, in Mudgee, Australia.
Griffin, P. 2001. *The profiles in practice: School reporting software.* Portsmouth, NH: Heinemann.
Griffin, P. 2007. The comfort of competence and the uncertainty of assessment. *Studies in Educational Evaluation* 33, no. 1: 87–99.
Griffin, P., and L. Burrill. 1995. Teachers' judgement as a major technique in literacy performance assessment: An international study. Paper presented at the annual meeting of the American Educational Research Association, April, in San Francisco.
Griffin, P., L. Murray, E. Care, A. Thomas, and P. Perri. 2010. Developmental assessment: Lifting literacy through professional learning teams. *Assessment in Education: Principles, Policy & Practice* 17, no. 4: 383–97.
Griffin, P., P.G. Smith, and N. Ridge. 2001. *The literacy profiles in practice: Toward authentic assessment.* Portsmouth, NH: Heinemann.
Gronlund, N.E. 1998. *Assessment of student achievement.* 6th ed. Boston, MA: Allyn & Bacon.

Holdaway, D. 1979. *The foundations of literacy.* Sydney: Ashton Scholastic.
Inhelder, B., and J. Piaget. 1958. *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures.* Trans. A. Parsons and S. Milgram. New York, NY: Basic Books.
Jaeger, R.M. 1995. Setting performance standards through two-stage judgemental policy capturing. *Applied Measurement in Education* 8, no. 1: 15–40.
Jenkinson, J.C. 1996. Identifying intellectual disability: Some problems in the measurement of intelligence and adaptive behaviour. *Australian Psychologist* 31, no. 2: 97–102.
Leinhardt, G. 1983. Novice and expert knowledge of individual student's achievement. *Educational Psychologist* 18, no. 3: 165–79.
Masters, G. 1982. A Rasch model for partial credit scoring. *Psychometrikca* 47, no. 2: 149–74.
McCloskey, G. 1990. Selecting and using the early childhood rating scales. *Topics in Early Childhood Special Education* 10, no. 3: 39–64.
McCutchen, D., and V.W. Berninger. 1999. Those who know, teach well: Helping teachers master literacy-related subject-matter knowledge. *Learning Disabilities Research & Practice* 14, no. 4: 215–26.
McDaniel, E. 1994. *Understanding educational measurement.* Madison, WI: WCB Brown and Benchmark.
McDonald, R.P. 1999. *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum Associates.
Messick, S. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist* 50, no. 9: 741–9.
Neisworth, J., and S. Bagnato. 1988. Assessment in early childhood special education: A typology of dependent measures. In *Early intervention for infants and children with handicaps*, ed. S. Odom and M. Karnes, 23–49. York, PA: Paul H. Brookes.
Piaget, J. 1947/2001. *The psychology of intelligence.* London: Routledge Classics.
Rasch, G. 1960/1980. *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedagogiske Institut.
Roberts, E., and P. Griffin. 2009. Profiling transitions in emotional development for students with additional learning needs. *Australasian Journal of Special Education* 33, no. 2: 151–60.
Silverstein, A.B., L. Brownlee, G. Legutki, and D.L. MacMillan. 1983. Convergent and discriminant validation of two methods of assessing three academic traits. *Journal of Special Education* 17, no. 1: 63–8.
Spear-Swerling, L., and R.J. Sternberg. 1996. *Off track: When poor readers become learning disabled.* Boulder, CO: Westview Press.
Victorian Auditor-General. 2007. *Program for students with disabilities: Program accountability.* PP No. 37, Session 2007–2008. Melbourne: Victorian Government Printer.
Victorian Curriculum and Assessment Authority (VCAA). 2006. Victorian Essential Learning Standards. http://vels.vcaa.vic.edu.au/ (accessed March 2, 2008).
Victorian Department of Education and Early Childhood Development (DEECD). 2009. Towards level one of the Victorian Essential Learning Standards: Curriculum advice. http://www.education.vic.gov.au/studentlearning/teachingresources/velslvl1.htm (accessed July 8, 2010).
Victorian Department of Education and Early Childhood Development (DEECD). 2011. Abilities Based Learning and Education Support. http://www.education.vic.gov.au/health-wellbeing/wellbeing/ables.htm (accessed November 10, 2011).
Vygotsky, L. 1929/1993. *The collected works of L.S. Vygotsky. Vol. 2: The fundamentals of defectology (abnormal psychology and learning disabilities).* Trans. R.W. Rieber and A. S. Carton. New York, NY: Plenum Press.
Wilson, M. 2005. *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum Associates.
Wolfe, E.W., and E.V. Smith, Jr. 2007. Instrument development tools and activities for measure validation using Rasch models: Part II – Validation activities. In *Rasch measurement: Advanced and specialized applications*, ed. E.V. Smith, Jr. and R.M. Smith, 243–90. Maple Grove, MN: JAM Press.

Woods, K. 2010. The design and validation of measures of communication and literacy to support the instruction of students with learning disabilities. Unpublished doctoral thesis, The University of Melbourne, Australia.

Woods, K., and P. Griffin. 2010. Teachers' use of developmental assessment to support communication proficiency for students with additional needs. Paper presented at the AARE International Research in Education Conference, November 28–December 2, in Melbourne, Australia.

Wright, B.D., J.M. Linacre, J.-E. Gustafson, and P. Martin-Lof. 1994. Reasonable mean-square fit values. *Rasch Measurement Transactions* 8, no. 3: 370. http://www.rasch.org/rmt/rmt83b.htm (accessed September 21, 2009).

Wright, B.D., and M.H. Stone. 1999. *Measurement essentials*. 2nd ed. Wilmington, DE: Wide Range.

Wu, M., R. Adams, and S. Haldane. 2008. *ConQuest*. Melbourne: Australian Council for Educational Research.

Appendix 1. Matrix of example draft items and hypothesised difficulty of their quality criteria as agreed by subject-matter experts

Draft quality criteria							
		Combines and orders symbols to convey messages and ideas	Plans and adapts ordering of ideas in written work, to express and explain ideas			Seeks out meaning of unfamiliar words	
	Interprets ordered sequences of words, symbols or signs to understand messages		Sequences symbols/ words to provide background information and convey meaning				Identifies sentences and some punctuation marks in text
	Identifies and names symbols and words	Uses and combines pictures or symbols to send simple messages		Interprets ordered sequences of photographs and/or pictures to understand messages			
					Tracks print from left to right across a page, and from top of page to bottom	'Reads' back own attempts at writing	Identifies spaces, letters and words in text
	Matches a symbol to an object, animal or person	Uses and produces pictures or symbols to represent objects or people		Identifies and names photographs and pictures	Turns pages of reading materials one at a time and from front to back	Follows or points to a line of text as it is read	Indicates the start and end of reading materials
			Matches and sorts pictures, photographs	Looks and points at realistic photographs		Asks for a familiar	

(*Continued*)

Appendix 1. (*Continued.*)

			and objects and presents them as they occur to him/her	of objects and may ask for the name of the object	Holds reading materials the right way up	story to be re-read	
	Looks at symbols and/or points to them	Responds to realistic photographs of familiar objects/ people	Makes choices between objects (or photographs of objects)	Responds to realistic photographs of familiar objects by looking, touching	Uses books and other reading materials for sensory/ exploratory activities		
Behavioural indicators (items)	*Responding to symbols*	*Matching symbols and meaning*	*Ordering symbols to express ideas*	*Responding to photographs/pictures*	*Knowing how to handle reading materials*	*Recognising that print has a consistent meaning*	*Using the terminology of printed text*