Screening of phonological awareness in the early elementary grades: an IRT approach

Judith M. T. Vloedgraven · Ludo Verhoeven

Abstract The purpose of the present study was to explore the possibilities for the assessment of growth in phonological awareness of children in kindergarten and first grade. Phonological awareness was measured using four sets of items involving rhyming, phoneme identification, phoneme blending, and phoneme segmentation. The results of an exploratory factor analysis and analyses conducted within the framework of item response theory showed one latent ability to underlie the different sets of items, which nevertheless differed in difficulty. Analyses in terms of the children's ability further showed the phonological awareness measures to be sensitive to growth. The amount of information supplied by the different sets of items depended on the children's level of ability. The conclusion that it is possible to accurately monitor the development of children's phonological awareness in the early elementary grades appears to be justified, and this possibility opens up new perspectives for the early screening for reading problems and dyslexia.

Keywords Phonological awareness · Item response theory · Dimensionality · Growth

During the last few decades, considerable attention has been paid by researchers, educators, and politicians to early screening for reading problems. Several studies have shown that early intervention can prevent later reading problems (National Institute of Child Health and Human Development, 2000; Slavin, Karweit, & Wasik, 1994). A major obstacle to early screening, however, is that the children have yet to receive formal literacy instruction (Fawcett & Nicolson, 2000). The adoption of a "predictor approach" is therefore called for and the precursors of reading acquisition must be identified.

J. M. T. Vloedgraven (✉) · L. Verhoeven
Department of Special Education, Radboud University Nijmegen,
P.O. Box 9104, 6500 HE Nijmegen, The Netherlands
e-mail: J.Vloedgraven@pwo.ru.nl

L. Verhoeven
e-mail: L.Verhoeven@pwo.ru.nl

According to relevant research literature, one of the strongest predictors of reading skills is phonological awareness. Stanovich (1994) and Elbro (1996) have both suggested that phonological awareness may be even more important than intelligence, vocabulary, and listening comprehension for the prediction of reading development. Phonological awareness refers to access to and an understanding of the sound structure of a spoken language, that is, the awareness that oral language can be broken down into individual words and, in turn, words into individual phonemes (cf. Wagner et al., 1997). Previous research has shown phonological awareness to be strongly related to early reading skills (Bradley & Bryant, 1983; Høien, Lundberg, Stanovich, & Bjaalid, 1995; Liberman, 1973; Perfetti, Beck, Bell, & Hughes, 1987; Wagner & Torgesen, 1987; Wagner, Torgesen, & Rashotte, 1994). There is also evidence that phonological deficits are the critical factor underlying reading problems (Elbro, Nielsen, & Petersen, 1994; Rack, Snowling, & Olson, 1992; Vellutino, Fletcher, Snowling, & Scanlon, 2004). In addition, interventions aimed at the improvement of phonological awareness have been shown to effectively promote learning to read (Lundberg, Frost, & Petersen, 1988; National Institute of Child Health and Human Development, 2000).

Problems with the measurement of phonological awareness

Researchers have encountered several problems with the measurement of phonological awareness. The first problem concerns the content validity of phonological awareness as a theoretical construct. A wide variety of tasks have been used to measure phonological awareness: rhyming tasks, phoneme counting tasks, sound comparison tasks, blending tasks, segmentation tasks, and deletion tasks. There is, however, ample evidence that these tasks differ in difficulty: rhyming tasks appear to be the easiest while tasks that require the manipulation of phonemes appear to be the most difficult (Adams, 1990; Chard & Dickson, 1999; Stanovich, Cunningham, & Cramer, 1984; Yopp, 1988). Just how these tasks relate to each other is far from clear. According to some researchers, the various aspects of phonological awareness measured by the tasks may actually reflect a single latent ability (Anthony & Francis, 2005; Anthony & Lonigan, 2004; Anthony et al., 2002; Stahl & Murray, 1994; Stanovich et al., 1984). In contrast, Yopp (1988) has argued that the construct of phonemic awareness consists of two highly related factors that nevertheless differ in the number of cognitive operations that they require: a simple phonemic awareness factor, which requires one operation, and a compound phonemic awareness factor, which requires an extra operation while holding the results of the first operation in memory. Muter, Hulme, Snowling, and Taylor (1997) have provided evidence for two other distinct factors and shown rhyming ability to be separate from segmentation ability. Høien et al. (1995) found three basic components to characterize phonological awareness: a phoneme factor, a syllable factor, and a rhyming factor. Clearly, there is still no consensus on the structure of phonological awareness.

Another problem with the measurement of phonological awareness is that the measures are often inaccurate. Inaccuracy problems may be caused by the fact that the suitability of a specific task depends on the child's level of development (Anthony & Lonigan, 2004; Chard & Dickson, 1999; Schatschneider, Francis, Foorman, Fletcher, & Mehta, 1999). According to Hambleton, Swaminathan and Rogers (1991), standard errors are only small when the difficulty of a test fits the ability of the examinee. Phonological awareness skills appear to lie along a continuum of increasing difficulty. By the end of kindergarten, for example, children have generally developed the ability to rhyme (Chard & Dickson, 1999).

If a rhyming task is then administered in first grade, most of the children will obtain a maximum score and, in this case, the exact level of each child's phonological awareness is still unknown. Conversely, a phonological awareness task may be too difficult at times. In a study by de Jong and van der Leij (1999), for example, no evidence was found for a relation between phonological awareness in kindergarten and reading performance in first grade due to the fact that two of the three tasks appeared to be too difficult for the kindergartners. These examples illustrate that if a task is not administered at the proper moment in a child's development, inaccurate measurement will be the result.

A related problem with the measurement of phonological awareness is that growth in this ability is hard to establish. If the children's abilities are not accurately measured, growth also cannot be accurately assessed. One possible solution to this problem is to administer different tasks of phonological awareness at different points in time (i.e., different developmental levels). However, comparison of task scores is made difficult, if not impossible, by the use of different scales and no demonstration of functional relations between the different scales. It is thus difficult to measure growth in phonological awareness, and this is a major problem for the identification of children who are at risk for reading problems and/or dyslexia. Several studies have shown that the measurement of phonological awareness and growth in this capacity are critical for the early identification of reading problems (Byrne, Fielding-Barnsley, & Ashley, 2000; Hindson et al., 2005; Spector, 1992). Growth in phonological awareness appeared to account for variance in reading in addition to that accounted for by the actual level of phonological awareness ability. Not only children's reading abilities but also their phonological abilities should thus be monitored during the development of beginning literacy as only the proper monitoring of children's (pre)literacy skills can enable the early identification of reading problems and dyslexia (Vellutino, Scanlon, & Lyon, 2000; Vellutino et al., 2004).

In sum, there are some major problems with the measurement of phonological awareness. First, it is unclear how the tasks used to measure the different aspects of phonological awareness relate to each other. Second, inaccurate measurement is a problem. And third, it is hard to measure growth. Most of the relevant studies use models from classical test theory (CTT) to assess the level of phonological awareness and predict the acquisition of beginning reading ability. However, the problems just described are difficult to resolve within the framework of CTT. A first problem with CTT is that scores have been found to depend on the particular set of items administered (i.e., be test-dependent). Another problem is that item parameters are group-dependent (i.e., characteristics such as item difficulty and discriminatory capacity appear to depend on the group to which the items are administered). Once again, these problems make it difficult to compare scores from different tasks (Hambleton & Jones, 1993). Even if the same task is completed by the child on different occasions (i.e., points in development), score comparison is still difficult because the accuracy of the measurement can vary across time.

An alternative approach is item response theory (IRT) or what is also known as latent trait theory (Hambleton et al., 1991). The distinctive feature of IRT models is that they relate item responses to ability: the difficulty of the items and the ability of persons are scaled on the same metric. Two assumptions hold for the specification of IRT models. First, it is assumed that the ability to be measured is unidimensional. Second, it is assumed that the relation between the latent trait and the probability of a correct response on a particular item can be described by the item characteristic curve (ICC). This curve is defined by one or more parameters, which determine the exact shape of the ICC. IRT has several advantages over CTT. A first advantage is that the estimated ability is test-independent, provided the

different tasks are constructed from an IRT-calibrated item bank. A second advantage is that the item parameter estimates are independent of the sample from which they are obtained. These two advantages make it possible to compare scores from different tasks. Another major advantage of IRT is the possibility to show the contribution of particular items and tasks to the assessment of ability (Lord, 1977). For the construction of early screening tasks, then, the test designer can select those items that provide the most information with regard to a particular ability and thereby develop the most accurate measures.

The present study

In the previous sections, the importance of measuring growth in phonological awareness was highlighted. Two methodological points appeared to be of particular importance. First, the construct of phonological awareness, as measured by various sets of items, has to be unidimensional. Second, the measures used to monitor the development of phonological awareness need to be sufficiently sensitive to growth (Kaminski & Good, 1996). The present study attempts to answer two questions related to these two methodological points. The first question is whether the different sets of items intended to measure phonological awareness appear to reflect a single underlying latent ability or several related abilities. For this purpose, the dimensionality of phonological awareness will be addressed from an IRT perspective. The second question is whether the items intended to measure phonological awareness can be used to measure growth from kindergarten through first grade. As already mentioned, the use of inaccurate measures is a major problem for the assessment of phonological awareness.

With regard to the first question, an initial attempt to identify the underlying structure of children's phonological awareness by the use of IRT was already undertaken by Schatschneider et al. (1999). The results of a factor analysis and the fit of an IRT model suggested that phonological awareness can be conceived as a unitary construct. A limitation on the study by Schatschneider et al., however, is that a rhyming task was not included. This means that the authors could neither confirm nor reject the findings of Muter et al. (1997) who found evidence suggesting that rhyming ability and segmentation ability may be separate. In the present study, we therefore administered four different types of items, which included rhyming items, to examine the underlying structure of phonological awareness. Schatschneider et al. also tested children speaking English while in the present study children speaking Dutch participated. It is a question whether the nature of phonological awareness is expected to be different in these two languages. Given the great overlap in phonological principles, it can be hypothesized that the sequence of phonological awareness development (i.e., from large units of sound to small units of sound) is the same for languages like English and Dutch (Ziegler & Goswami, 2005). Because of the fact that the relation between phonological awareness and reading is bidirectional (cf. Perfetti et al., 1987), it is important to look at the differences between the orthographies of the two languages as well. Seymour, Aro, and Erskine (2003) concluded on the basis of a crosslinguistic comparison of different orthographies that Dutch and English orthography share a complex syllabic structure, but differ in orthographic depth. Because the orthographic depth in Dutch is evaluated to be smaller than in English, we expect Dutch children to be faster in developing phonological awareness without a change in the underlying structure of phonological awareness. Educational environment neither seems to alter the structure of phonological awareness because both in the Netherlands and in England a phonics teaching method is primarily used. Given the fact that the orthographic

depth only seems to influence the rate of development in phonological awareness and not the underlying structure, the present study can by and large be seen as a replication of the study by Schatschneider et al. with, as central hypothesis, a unidimensional structure of phonological awareness.

Related to the issue of the unidimensional or multidimensional structure of phonological awareness is the issue of the relative difficulty of the different sets of items. As already noted, the various sets of items used to measure phonological awareness have been found to differ in difficulty. However, the exact differences between the various item sets are still open to investigation. In addition to these differences between tasks, differences in linguistic complexity within tasks appear to influence phonological awareness (Anthony & Francis, 2005). For example, according to Schreuder and van Bon (1989) the consonant–vowel (CV) structure is an important determinant. Therefore, as a next step, we have investigated the differences in difficulty of various CV structures.

The second question to be investigated is whether the items, measuring phonological awareness, are able to measure growth from kindergarten to grade 1. As mentioned earlier, the lack of accurate measures for the assessment of growth in phonological awareness is a major problem. If the measures used in the present study appear to be sensitive to growth in phonological awareness, then the accuracy for the different sets of items and the complete set of items will be examined for a range of ability scores. Results will show which set(s) of items are of importance in assessing the ability of kindergartners and first graders.

Method

Participants

A total of 172 children in their second year of kindergarten (88 boys and 84 girls) and 173 first-grade children (89 boys and 84 girls) were randomly selected from 12 elementary schools in the east part of The Netherlands. All of the children spoke Dutch and were from a variety of socioeconomic backgrounds. In the Dutch educational system, children visit school from the moment they are 4 years old, after which they spend 2 years in kindergarten. After these 2 years, children enter first grade. In kindergarten, literacy education is generally limited to some language games to stimulate phonological awareness and beginning literacy. Formal instruction in reading and spelling instruction starts in first grade and from that moment, explicit instruction in phonics is offered.

The children were tested in April or May of 2005. At the time of testing, the mean age of the kindergartners was 6 years and 1 month (SD=4.4 months); the mean age of the first graders was 7 years and 1 month (SD=3.9 months).

Materials

To select tasks for phonological awareness, we have looked at the extent to which a task represents phonological awareness ability. In addition, the predictive value for reading performance was taken into account. Furthermore, we have selected various tasks that according to the literature are known to differ in difficulty to be able to accurately assess the ability of both high- and low-ability individuals. Taking these criteria into consideration, the following four tasks were selected: rhyming, phoneme identification, phoneme blending, and phoneme segmentation (Adams, 1990; Chard & Dickson, 1999; Høien et al., 1995; Vellutino & Scanlon, 1987; Yopp, 1988). The tasks consisted of high-frequency

monosyllabic words containing two, three, four, or five phonemes. The CV structure of the target words varied. The target words were selected from current Dutch word frequency list (Schaerlaekens, Kohnstamm, & Lejaegere, 1999; Schrooten & Vermeer, 1994). Given that a developmen-tally appropriate test for phonological awareness should not overload working memory (Reitsma, 2003), all of the words were presented both auditorily and visually. In all cases, the presented pictures were previously named to be certain that the correct names were associated with the pictures.

Rhyming

Three pictures were shown to the children. The target word was then presented auditorily (via the computer) and the children were asked to select the word that rhymed with the target word. Distractors were chosen so that they contrast much with the target word. All of the target words were consonant–vowel–consonant (CVC) words. Thirty items were administered to the children.

Phoneme identification

Three pictures were presented to the children. The target phoneme was then pronounced along with a word that started with the same phoneme. The child's task was to select the picture that started with the same sound that the target word started with. Only consonants were used as target phonemes and articulated as sounds. All of the target words were CVC words. Thirty items were administered to the children.

Phoneme blending

Three pictures were presented to the children. The isolated phonemes from the target word were then pronounced in their correct order. The child's task was to select the picture that represented the target word. To be able to measure blending ability as pure as possible, distractors contained one or more of the phonemes of the target word. The target words consisted of three, four, or five phonemes with different CV structures. Each child was given 40 items.

Phoneme segmentation

The target word was presented visually and auditorily. The child was asked to say the phonemes of the target word separately in the correct order. Word length was two, three, four, or five phonemes. Each child was given 40 items.

Procedure

All of the tasks were administered individually and presented on a computer. For kin-dergartners, the tasks were administered in two sessions of about 20 min each because of their relatively short attention spans. First graders were tested in one session, which took approximately 20 min to complete. The rhyming task was presented to the kindergartners only because it is well known that this task is the easiest phonological awareness task and most suitable for kindergartners (Adams, 1990; Chard & Dickson, 1999). In addition, Schatschneider et al. (1999) has shown identification of the first sound in a word to provide a poor estimate of phonological ability for first-grade children because the task is too easy for this age. Therefore, this task was also administered in kindergarten only.

Booklet	N	Grade	rhyming			identification			blending			segmentation		
			A	B	C	D	E	F	G	H	I	J	K	L
1	60	kindergarten	x	x		x	x		x	x		x	x	
2	55	kindergarten		x	x		x	x		x	x		x	x
3	57	kindergarten	x		x	x		x	x		x	x		x
4	57	first grade							x	x		x	x	
5	59	first grade								x	x		x	x
6	57	first grade							x		x	x		x

Fig. 1 Anchor test design for rhyming, phoneme identification, phoneme blending, and phoneme segmentation

Three practice items preceded each task to familiarize the children with the testing procedure. After each practice item, the experimenter provided feedback on the correctness of the child's response. If the child gave an incorrect answer, the correct answer was provided.

As mentioned earlier, phonological awareness was measured by four different tasks. An item bank has been constructed that contained four sets of items representing these four different tasks (i.e., 45 rhyming items, 45 phoneme identification items, 60 phoneme blending items, and 60 phoneme segmentation items). Because it was not feasible to present all of the items to all of the children, we used a structural incomplete design, called the anchor test design (Petersen, Kolen, & Hoover, 1989). Therefore, all of the items of a task were divided in three modules and each child was given two of the three modules of a task (i.e., booklet). To be able to administer all of the items from the item bank, different groups of children were given different booklets (i.e., different combinations of modules). Characteristic of this design is the link between booklets: the different booklets have certain anchor items in common. And by the use of these anchor items, it is possible to present all of the items at one scale of measurement. The design of the study is presented in Fig. 1.

Statistical analyses

The four sets of items intended to measure phonological awareness were submitted to several analyses to establish their underlying structure. As a first step, we performed an exploratory factor analysis on the matrix with tetrachoric correlations of the items. The factor analysis was conducted using the minimum residuals (MINRES) method (Harman & Jones, 1966). The MINRES method minimizes the sum of squared residuals, resulting in a matrix of factor loadings.

The next step in the analyses involved the use of IRT models. IRT models diverge depending on whether the relation between item performance and knowledge is taken as a one-, two-, or three-parameter logistic function. In the one-parameter model only difficulty parameters are estimated; in the two-parameter model difficulty and discrimination parameters are estimated; and three-parameter models take into account the effect of item guessing, difficulty, and discrimination. A main advantage of the one-parameter model is the possibility to apply the conditional maximum likelihood (CML) procedure to estimate item parameters and the sampling independence implied by it (Verhelst & Glas, 1994). In contrast, the use of CML is impossible in the two other models. However, a drawback of the one-parameter model is that it is not very realistic that discrimination indices are the same for all of the items. Therefore, we used the one-parameter logistic model (OPLM), which is a synthesis of the one- and two-parameter model. The most important feature of

the OPLM is that difficulty parameters are estimated and discrimination parameters are dealt with as known constants (i.e., discrimination indices can vary, but have discrete values). The discrimination indices are based on a geometric mean of 3. The difficulty parameter provides information on the difficulty of an item and is the point on the ability scale where the probability of a correct answer is 0.5. The discrimination parameter refers to the slope of the curve at its steepest position. This parameter indicates how well an item discriminates between high- and low-ability individuals. To estimate the item parameters, we used the CML procedure. A one-way ANOVA was conducted, followed by a Tukey's test, to establish the differences between the difficulty parameters for the four sets of items. Person parameters (ability) were estimated by means of the weighted maximum likelihood (WML) procedure. To assess the ability distributions of the two populations, the expected a posteriori estimation was used because the use of the WML procedure for this purpose can lead to an overestimation of means and standard deviations. Thereafter, the Cohen's d was used to measure the strength of growth from kindergarten through first grade.

Results

Underlying structure of phonological awareness

First, a matrix with tetrachoric correlations was computed for all of the items. Next, a two-factor analysis on this matrix of correlations using MINRES was conducted. On the basis of the factor loadings for the items on both factors, the eigenvalues were then computed. This resulted in a powerful first factor, which extracted 82% of the total variance. The second factor accounted for 18% of the total variance. It should be noted that tetrachoric correlations have relatively large standard errors (Brown, 1977). In the case of a small sample size, this complicates the identification of the correct number of factors. The results of the analyses with MINRES should therefore be interpreted with caution. The large percentage of the variance explained by the first factor and the significant difference in the contribution of the second factor can nevertheless be seen as evidence for unidimensionality.

Assuming unidimensionality, we examined whether the OPLM fits the data. To assess model fit, both the item-oriented statistics and an overall statistic were computed. First, the OPLM can be used to determine if the individual items fit the same latent trait. For each item, an indication of the fit into the model is provided by the p value. For 14 of the 210 items, a misfit was detected and these items were then deleted. A formal means to judge the distribution of the p values for all of the items is not available (Verhelst, Glas, & Verstralen, 1995). However, it is certain that a majority of low p values indicate model violations, and it is desirable that the frequencies of the p values be about equal for each interval (i.e., 0–0.10,…, 0.90–1.00). It appeared that none of the items had a p value lower than 0.05. Moreover, the distribution of the frequencies was fairly balanced across the intervals, showing model fit. Second, additional information about the model fit can be provided by the overall R1c test. The R1c value was 584.01 ($df=537$, $p=0.08$), which suggests that the different sets of items can be included in the same scale.

When the OPLM fits the data, the invariance of the ability and item parameters can be established. Ability invariance means that the estimated ability of each person does not depend on the specific set of items administered. Invariance of item statistics means that the item parameters derived from the model are independent of the specific sample (Hambleton et al., 1991). To assess whether the OPLM fits the data, we investigated these two
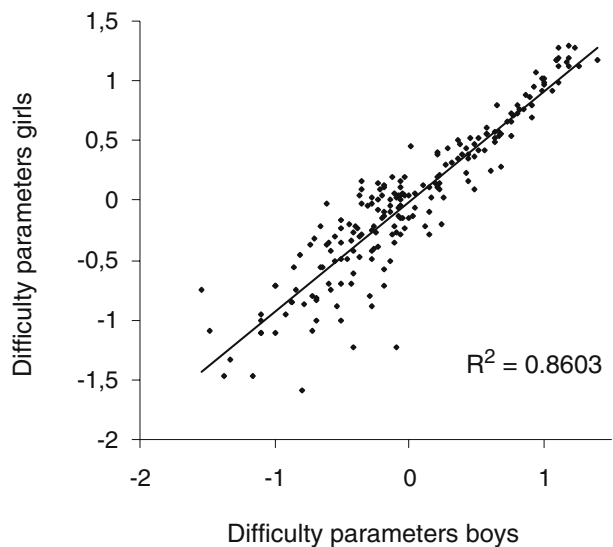
Fig. 2 Invariance of ability parameters
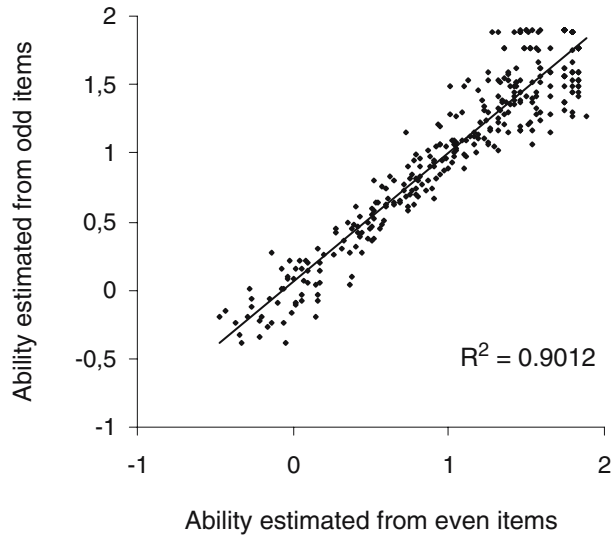
properties. First, we dealt with the question of ability invariance. The items were divided into two subsets: even and odd items. For each child, the ability parameters were then estimated for the two subsets of items. Thereafter, a scatterplot of the pairs of ability estimates was made (see Fig. 2). If the ability estimates are invariant, the plot should demonstrate a straight line. As can be seen, a strong linear association was indeed found to hold between the ability estimates for the even and odd items.

The invariance of the item statistics was next examined by determining the associations between the difficulty parameters estimated from two different samples. For this purpose, the total sample was split into two subpopulations: boys and girls. The estimated difficulty parameters for the boys were then plotted against the estimated difficulty parameters for the

Fig. 3 Invariance of difficulty parameters

girls. The corresponding scatter plot is presented in Fig. 3. As can be seen, the relation between the difficulty parameters for the two different samples appears to be linear, which clearly indicates the invariance of the item parameters. And these results provide strong evidence for the assumption of a unidimensional underlying structure for phonological awareness.

Item parameters

The preceding results showed phonological awareness to be well represented by a single underlying scale. The next question is just how the various sets of items measuring phonological awareness relate in terms of difficulty. In Table 1, the average difficulty and discrimination parameters for the four sets of items measuring phonological awareness are presented from least to most difficult. The rhyming items turned out to be the easiest, and the phoneme segmentation items turned out to be the most difficult. The items measuring phoneme blending and phoneme identification occurred in between. A one-way ANOVA showed the differences in difficulty to also be significant $[F(3, 192)=188.78, p<0.01]$. Multiple comparisons were next conducted using the Tukey procedure to determine which pairs of item sets differed significantly from each other. The analyses showed all of the pairs of item sets to differ significantly, with the exception of the difference between the phoneme identification items and the phoneme blending items. As can be seen in Table 1, the various sets of items also differ in their capacity to discriminate between high- and low-ability individuals. Items measuring phoneme segmentation turned out to be most discriminating while items measuring rhyming ability turned out to be least discriminating.

In addition to differences between the four sets of items, effects of CV structure within tasks have been investigated. With respect to phoneme blending, items were divided into three sets of items: (1) CVC; (2) CVCC and CCVC; and (3) CCVCC, CCCVC, and CVCCC. Items of the first item set appeared to be the easiest and items of the third item set appeared to be the most difficult. However, a one-way ANOVA showed the differences in difficulty not to be significant $[F(2, 52)=2.17, p=0.124]$. With respect to phoneme segmentation, items were divided into five item sets: (1) CV and VC; (2) CVC; (3) CCV and VCC; (4) CCVC and CVCC; and (5) CCVCC, CCCVC, and CVCCC. Table 2 shows the average difficulty parameters for the various CV structures and ranks the different item sets from least to most difficult. Results of an ANOVA analysis revealed that differences in difficulty between the various CV structures were significant $[F(4, 49)=29.16, p<.01]$. This leads to the next question, that is, between which pairs of item sets are the differences significant? Results of Tukey's test showed most of the pairs to differ significantly, except for the distinction between CV, VC, and CVC; between CCV, VCC and CCVC, CVCC; and between CCVC, CVCC and CCVCC, CCCVC, CVCCC. Considering these findings,

Table 1 Average difficulty and discrimination parameters for the four sets of items measuring phonological awareness

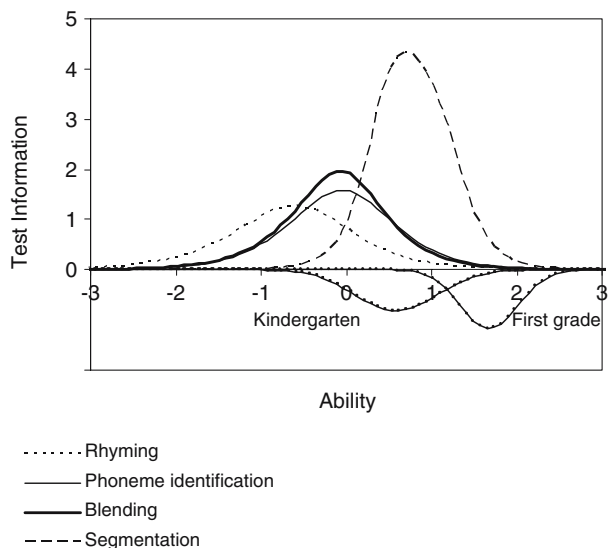| Set of items	Difficulty parameter		Discrimination parameter	
	M	SD	M	SD
Rhyming	−0.69	0.36	2.33	0.68
Phoneme blending	−0.16	0.32	2.85	0.78
Phoneme identification	−0.08	0.25	2.57	0.63
Phoneme segmentation	0.77	0.30	4.98	1.04

Table 2 Average difficulty parameters for the different CV structures within phoneme segmentation

Set of items	Difficulty parameter	
	M	SD
CV, VC	0.37	0.13
CVC	0.56	0.15
CCV, VCC	0.81	0.23
CCVC, CVCC	0.97	0.20
CCVCC, CCCVC, CVCCC	1.10	0.13

we may conclude that merely lengthening a word does not have an effect on the difficulty of segmentation. Differences are only significant when a pair of items differs in the distribution of consonant clusters. Another finding that confirms the effect of consonant clusters is that the difference between CVC and CCV, VCC words turns out to be significant, with the latter being the most difficult, despite the fact that they are similar in word length. The existence of one or more consonant clusters in a word thus appears to complicate the performance in a segmentation task.

Growth in phonological awareness

To establish whether the phonological awareness measures are sensitive to growth, the progress of the children from kindergarten to first grade was investigated. The results concerning growth should be interpreted cautiously because the subjects are separate groups of kindergartners and first graders. Despite this, results are expected to give valuable indications about the development in phonological awareness. The ability distributions for the kindergartners ($M=0.582$, SD=0.485) and first graders ($M=1.677$, SD=0.341) are presented in Fig. 4 below the *x*-axis. As can be seen, the first graders improved importantly. The Cohen's *d*, which is an objective measure of the strength of growth, was 2.60 (Cohen, 1988). Given that an effect of 0.80 is interpreted as a large effect, an effect of 2.60 can be judged to be a substantial effect.

Fig. 4 Test information functions for the four sets of items measuring phonological awareness

Information functions and accuracy of ability estimates

The difficulty and discrimination parameters for the different sets of items used to measure phonological awareness were just compared. A restriction on these comparisons is that they lack information on which task or tasks may be most useful for the population of interest. As already mentioned, one of the advantages of IRT is the possibility to show the contributions of particular items and sets of items. This can be realized by calculating test information functions. These functions link information from both the difficulty and discrimination parameters. In such a manner, it is possible to specify just how well a task estimates ability across the total distribution of ability. In other words, information functions indicate the accuracy of measurement of the different tasks for different ability levels. Information functions frequently diverge across the range of ability scores; a task is possibly more informative for high-scoring individuals than for low-scoring individuals or the other way round. The analogous measure for information functions in CTT is reliability. Nevertheless, reliability in IRT cannot be compared with reliability in CTT without any problems. Because in IRT reliability is different for each point of the latent trait scale, it is also called local reliability. However, it is possible to transform the local measurement precision in IRT to the classical measure of reliability. Before we continue with accuracy of measurement in IRT, first we will mention the classical indices for reliability because these indices are easier to interpret. The reliability of each task from a CTT framework was determined using the MAcc coefficient (Verhelst et al., 1995). For rhyming the MAcc appeared to be 0.83, for phoneme identification 0.91, for phoneme blending 0.96, and for phoneme segmentation 0.99. All of the tasks appeared to be sufficiently accurate in estimating phonological awareness skills.

As a next step, the information functions for each of the four sets of items were thus computed. Given that the amount of information provided by a set of items is influenced by the number of items, this number was taken into account. The four information functions are presented in Fig. 4. In the same figure, the ability distributions for both subpopulations (i.e., kindergartners and first graders) are plotted. This gives the opportunity to see at a glance which specific set of items is most informative at a particular level of ability; the higher the information function, the more accurate the estimates of ability are at a given point of the ability scale. The maximum of the information function for phoneme segmentation appears to be highest and occurs at approximately 0.75 on the ability scale,

Fig. 5 Total test information function
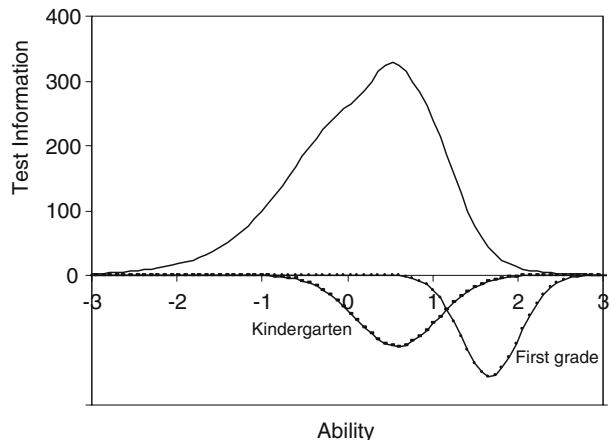
which shows the segmentation items to be most informative at that point of the ability scale. In addition, relative to the information functions for rhyming, phoneme identification, and phoneme blending, the information function for phoneme segmentation has moved to the right along the ability axis. This means that this task more accurately estimates the ability of kindergartners with a higher ability and of first graders. However, items measuring phoneme segmentation provide weaker estimates for those children at the highest end of the ability range. Although the items from the three other sets are generally less informative, they nevertheless provide more information about the relevant capacities of the kindergartners with lower levels of ability than the phoneme segmentation items.

In addition to the information functions for the different sets of items, the total test information function for all of the items was also calculated (see Fig. 5). As can be seen, the four sets of items together provide the most precise estimates of children with an ability score of about 0.5, which corresponds to the average ability score for kindergartners. The total test does not provide an accurate estimate of the ability of first graders with an average or above average ability score. However, the estimates for kindergartners at the lower end of the ability distribution are still satisfactory.

Conclusions and discussion

Several conclusions can be drawn on the basis of the present results. With respect to the unidimensionality of phonological awareness, an exploratory factor analysis showed one latent ability to underlie the different sets of items used to measure phonological awareness within the context of the present study. The results showed the first factor to account for a large percentage of the variance and an enormous difference in the contributions of the first and second factors, which are findings highly indicative of a single underlying factor (i.e., the unidimensionality of phonological awareness). The assumption of unidimensionality was further investigated using a model based on IRT. Both the item-oriented and overall statistical tests showed the OPLM to fit the data. Ability invariance and item parameter invariance were also demonstrated, which supports the conclusion that the various sets of items used to measure phonological awareness indeed reflect one and the same latent ability. This result is in accordance with the outcomes of the study by Schatschneider et al. (1999) who tested English-speaking children. As we expected, differences in orthographic depth between English and Dutch did not influence the underlying structure of phonological awareness. In contrast to Schatschneider et al., we included a rhyming task. The results of the present study further support Treiman's (1985) claim that although rhyming deals with larger linguistic units than phonemes, the cognitive operations needed to rhyme also require awareness of abstract speech representations.

Given the indications that one latent ability underlies the different sets of items, the next issue to be addressed was the relative difficulty of the different sets of items. The results of the ANOVA and Tukey analyses indeed showed the sets of items to differ in difficulty. The rhyming items appeared to be the easiest and the segmentation items appeared to be the most difficult with the phoneme blending and phoneme identification items occurring in between. The differences between all of the pairs of item sets were significant, with the exception of the difference between the sets of items of phoneme blending and phoneme identification. These results show the cognitive task requirements for the sets of items to clearly differ despite a single underlying ability. The finding that the rhyming items were the easiest is in agreement with the findings of many other studies (Adams, 1990; Chard & Dickson, 1999; Stanovich et al., 1984; Yopp, 1988). The present findings also confirm the

findings of previous research showing the extreme difficulty of phoneme segmentation and the intermediate difficulty of phoneme blending and phoneme identification (Høien et al., 1995).

Furthermore, our examination of the relative difficulty of various CV structures within tasks showed no effect for phoneme blending, which may be due to the relative ease of this task. However, differences in CV structure in the segmentation task appeared to be significant: the longer the word, the more difficult it was to segment that word in separate phonemes. Closer inspection of the significant differences between all pairs of items revealed that longer words were only harder to segment when one or more consonant clusters were added. This finding is in agreement with previous researches (Arnqvist, 1992; Schreuder & van Bon, 1989; Treiman & Weatherston, 1992).

The second issue addressed in the present study was whether the phonological awareness measures were also sensitive to growth. The strength of growth from kindergarten to first grade was indicated by Cohen's d, which appeared to be 2.60 and can thus be interpreted as a substantial effect. It is thus possible to measure growth in phonological awareness during the development of beginning literacy. The accuracy of the various sets of items across the spectrum of kindergartners and first graders with different degrees of ability was determined by investigating the information functions of the four sets of items and the ability distributions of both subgroups. On the basis of these results, the assumption that the appropriateness of a particular task depends on the level of child development (Anthony & Lonigan, 2004; Chard & Dickson, 1999; Schatschneider et al., 1999) received support. Our results indeed showed the usefulness of the various sets of items to depend upon the difficulty of the items and the abilities of the child. The IRT model showed the phoneme segmentation items to provide the most information about ability. Closer inspection of the information function showed the phoneme segmentation set of items to estimate the ability of higher scoring kindergartners and lower scoring first graders most accurately; for lower scoring kindergartners and higher scoring first graders, however, the estimates were less accurate. Although the information provided by rhyming performance, phoneme blending, and phoneme identification is relatively low, inclusion of these sets of items in addition to phoneme segmentation items in an instrument for early screening may be critical as exactly these aspects of phonological awareness appear to be most informative for those children at the lower end of the ability continuum.

Information functions are determined to a great extent by the discriminating power of the items. As mentioned earlier, segmentation items are the most discriminating. From the information function of segmentation, we can derive that this is especially valid for kindergartners and lower scoring first graders. This can be explained by the fact that segmentation items best suited the ability level of these children. A striking result was that the discriminating power of the phoneme segmentation items decreased substantially as children improve their phonological awareness ability during first grade. Due to the start of literacy education with explicit instruction in phonics in first grade, children generally master the ability to segment words into phonemes in the course of first grade. At the end of this year, segmentation items are too easy for most of the children and thus we were not able any more to differentiate between high- and low-ability children.

When the total test information for all of the items is examined, we can conclude that the most accurate estimates are obtained for the average kindergartner. The four sets of items together adequately measure the ability of lower scoring kindergartners. However, as the children's abilities increase during first grade, the ability estimates become less and less accurate. These results suggest that inclusion of the four sets of items in a screening

instrument can be recommended but that another set of more difficult items should also be included to improve the accuracy of measurement for first graders in particular. Adams (1990), for example, has described the different levels of difficulty for phonemic awareness and found phoneme manipulation to be most difficult as this requires the addition or omission of phonemes to formulate a new word.

In sum, the results of the present study have shown that it is possible to measure growth in phonological awareness. The various sets of items used to measure phonological awareness could be placed along a single ability scale and were found to measure changes in phonological awareness (i.e., growth). However, a refinement of the ability scale is necessary to attain more accurate ability estimates for the higher end of the ability range.

The findings of the present study have some important implications for the early screening of reading problems and dyslexia. The results show that the development of phonological awareness can be accurately monitored. As already stated, several studies showed growth in phonological awareness to add unique information to the prediction of reading, which highlights the importance of monitoring the development of phonological awareness. When McBride-Chang, Wagner, and Chang (1997) investigated the development of phonological awareness, they found evidence for Matthew effects (Stanovich, 1986) of prereading skills. That is, children who started with a higher level of phonological awareness tended to improve their level of skill more quickly than children who started with a lower level of phonological awareness. Future research within the framework of IRT will show whether these results can be confirmed. If such Matthew effects are indeed found for phonological awareness, then a successful early start can be seen to be paramount and the value of early screening for reading problems and dyslexia thus reinforced. However, we have seen that the discriminating power of phonological awareness tasks decreases in the course of first grade. It is important to note that, despite adding a more difficult task to try to improve the accuracy of the measures in first grade, the value of screening of phonological awareness for the prediction of reading will steadily decrease as children's abilities increase. This finding is in concordance with the conclusion of de Jong and van der Leij (1999) that the predictive value of phonological awareness tasks is limited to the early phases of learning to read in The Netherlands. As a consequence, for the early identification of reading problems and dyslexia, it is thus of major importance to assess children's phonological awareness in kindergarten and in the first half of first grade. And most important is the measurement of growth in phonological awareness. Like intervention in reading is used to help in the distinction between reading difficulties caused by cognitive deficits and those caused by instructional deficits (Vellutino et al., 1996), intervention in phonological awareness can aid in the same way. It is clear that phonological abilities should be monitored to be able to identify children who show a phonological deficit given the fact that, even after receiving intervention, they hardly improve.

As already mentioned, the prediction of future reading skills clearly depends on the accuracy with which prereading skills are measured. The results of the present study have confirmed earlier findings showing the precision of the measurement by a set of items to depend on the child's actual level of ability. Valuable information has been provided on the utility of particular sets of items for kindergartners and first graders with different levels of ability. These findings further suggest that the influence of various aspects of phonological awareness on later reading skill may be constrained, which has also been found by de Jong and van der Leij (1999). In contrast to other studies and due to inaccurate measures, de Jong and van der Leij could not demonstrate that phonological awareness in kindergarten was related to later reading. This issue certainly merits further study and longitudinal study

in particular to show which tasks best predict reading skill at different moments in a child's development. Greater research on this topic is also of major importance for improved early screening of children who are possibly at risk for reading failure.

The present study can be seen as a first attempt to investigate the underlying structure of phonological awareness as regards Dutch language from an IRT perspective. In a follow-up study, we will enlarge the sample size and also collect data about reading and spelling to be able to relate phonological awareness measures with reading and spelling measures.

Do the results of the present study also lead to recommendations for ongoing practice? On the one hand, we have demonstrated that the various sets of items measuring phonological awareness are represented by a single underlying ability. This means that practitioners can use the full scale as a sensitive screening instrument for phonological awareness. On the other hand, we have shown the cognitive task requirements of those sets of items to clearly differ. Given this finding, it is recommended that subscales be used to further diagnose the phonological awareness of children whose scores on the latent scale qualify them as at risk. Accordingly, specific phonological awareness training programs can then be implemented (Bus & van Ijzendoorn, 1999; Troia, 1999). More research is needed to solve this dilemma.

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Anthony, J. L., & Francis, D. J. (2005). Development of phonological awareness. *Current Directions in Psychological Science, 14*, 255–259.

Anthony, J. L., & Lonigan, C. J. (2004). The nature of phonological awareness: Converging evidence from four studies of preschool and early grade school children. *Journal of Educational Psychology, 96*, 43–55.

Anthony, J. L., Lonigan, C. J., Burgess, S. R., Driscoll Bacon, K., Phillips, B. M., & Cantor, B. G. (2002). Structure of preschool phonological sensitivity: Overlapping sensitivity to rhyme, words, syllables, and phonemes. *Journal of Experimental Child Psychology, 82*, 65–92.

Arnqvist, A. (1992). The impact of consonant clusters on preschool children's phonemic awareness: A comparison between readers and nonreaders. *Scandinavian Journal of Psychology, 33*, 29–35.

Bradley, L., & Bryant, P. E. (1983). Categorizing sounds and learning to read: A causal connection. *Nature, 301*, 419–421.

Byrne, B., Fielding-Barnsley, R., & Ashley, L. (2000). Effects of preschool phoneme identity training after six years: Outcome level distinguished from rate response. *Journal of Educational Psychology, 92*, 659–667.

Brown, M. B. (1977). Algorithm AS 116: The tetrachoric correlation and its standard error. *Applied Statistics, 26*, 343–351.

Bus, A. G., & van Ijzendoorn, M. H. (1999). Phonological awareness and early reading: A meta-analysis of experimental training studies. *Journal of Educational Psychology, 91*, 403–414.

Chard, D. J., & Dickson, S. V. (1999). Phonological awareness: Instructional and assessment guidelines. *Intervention in School and Clinic, 34*, 261–270.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

de Jong, P. F., & van der Leij, A. (1999). Specific contributions of phonological abilities to early reading acquisition: Results from a Dutch latent variable longitudinal study. *Journal of Educational Psychology, 91*, 450–476.

Elbro, C. (1996). Early linguistic abilities and reading development: A review and a hypothesis. *Reading and Writing: An Interdisciplinary Journal, 8*, 453–485.

Elbro, C., Nielsen, I., & Petersen, D. K. (1994). Dyslexia in adults: Evidence for deficits in non-word reading and in the phonological representations of lexical items. *Annals of Dyslexia, 44*, 205–226.

Fawcett, A. J., & Nicolson, R. I. (2000). Systematic screening and intervention for reading difficulty. In N. A. Badian (Ed.), *Prediction and prevention of reading failure* (pp. 57–85). Baltimore: York Press.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement, Issues and Practice, 12*, 38–47.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Harman, H. H., & Jones, W. H. (1966). Factor analysis by minimizing residuals (MINRES). *Psychometrika, 31*, 351–369.

Hindson, B. A., Byrne, B., Fielding-Barnsley, R., Newman, C., Hine, D. W., & Shankweiler, D. (2005). Assessment and early instruction of preschool children at risk for reading disability. *Journal of Educational Psychology, 97*, 687–704.

Høien, T., Lundberg, I., Stanovich, K. E., & Bjaalid, I. (1995). Components of phonological awareness. *Reading and Writing: An Interdisciplinary Journal, 7*, 171–188.

Kaminski, R. A., & Good, R. H. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review, 25*, 215–227.

Liberman, I. (1973). Segmentation of the spoken word and reading acquisition. *Bulletin of the Orton Society, 23*, 65–77.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*, 117–138.

Lundberg, I., Frost, J., & Petersen, O. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly, 23*, 263–284.

McBride-Chang, C., Wagner, R., & Chang, L. (1997). Growth modeling of phonological awareness. *Journal of Educational Psychology, 89*, 621–630.

Muter, V., Hulme, C., Snowling, M., & Taylor, S. (1997). Segmentation, not rhyming, predicts early progress in learning to read. *Journal of Experimental Child Psychology, 65*, 370–396.

National Institute of Child Health and Human Development (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH publication no. 00-4769). Washington, DC: US Government Printing Office.

Perfetti, C., Beck, I., Bell, L., & Hughes, C. (1987). Phonemic knowledge and learning to read are reciprocal: A longitudinal study of first grade children. *Merrill–Palmer Quarterly, 33*, 283–319.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York, USA: American Council on Education.

Rack, J. P., Snowling, M. J., & Olson, R. K. (1992). The nonword reading deficit in developmental dyslexia: A review. *Reading Research Quarterly, 27*, 28–53.

Reitsma, P. (2003). Precursors of phonemic awareness. In L. Verhoeven, C. Elbro, P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 33–48). Amsterdam: Philadelphia.

Schaerlaekens, A., Kohnstamm, D., & Lejaegere, M. (1999). *Streeflijst woordenschat voor zesjarigen* [Strived vocabulary list for six-year olds]. Lisse: Swets & Zeitlinger.

Schatschneider, C., Francis, D. J., Foorman, B. F., Fletcher, J. M., & Mehta, P. (1999). The dimensionality of phonological awareness: An application of item response theory. *Journal of Educational Psychology, 91*, 467–478.

Schreuder, R., & van Bon, W. H. J. (1989). Phonemic analysis: Effects of word properties. *Journal of Research in Reading, 12*, 59–78.

Schrooten, W., & Vermeer, A. (1994). Woorden in het basisonderwijs. 15.000 woorden aangeboden aan leerlingen [Words in primary education. 15,000 words offered to pupils]. *Studies in Meertaligheid, 6*, 1–44.

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*, 143–174.

Slavin, R. E., Karweit, N. L., & Wasik, B. A. (1994). *Preventing early school failure: Research, policy, and practice.* Boston: Allyn & Bacon.

Spector, J. E. (1992). Predicting progress in beginning reading: Dynamic assessment of phonemic awareness. *Journal of Educational Psychology, 84*, 353–363.

Stahl, S. A., & Murray, B. A. (1994). Defining phonological awareness and its relationship to early reading. *Journal of Educational Psychology, 86*, 221–234.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360–407.

Stanovich, K. E. (1994). Romance and reality. *Reading Teacher, 47*, 280–291.

Stanovich, K. E., Cunningham, A. E., & Cramer, B. B. (1984). Assessing phonological awareness in kindergarten children: Issues of task comparability. *Journal of Experimental Child Psychology, 38*, 175–190.

Troia, G. A. (1999). Phonological awareness intervention research: A critical review of the experimental methodology. *Reading Research Quarterly, 34*, 28–52.

Treiman, R. (1985). Onset and rimes as units of spoken syllables: Evidence from children. *Journal of Experimental Child Psychology, 39*, 161–181.

Treiman, R., & Weatherston, S. (1992). Effects of linguistic structure on children's ability to isolate initial consonants. *Journal of Educational Psychology, 84*, 174–181.

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): What we have learned in the past four decades? *Journal of Child Psychology and Psychiatry, 45*, 2–40.

Vellutino, F. R., & Scanlon, D. M. (1987). Phonological coding, phonological awareness and reading ability: Evidence from a longitudinal and experimental study. *Merrill–Palmer Quarterly, 33*, 321–363.

Vellutino, F. R., Scanlon, D. M., & Lyon, G. R. (2000). Differentiating between difficult-to-remediate and readily remediated poor readers: More evidence against the IQ-achievement discrepancy definition of reading disability. *Journal of Learning Disabilities, 33*, 223–238.

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S. G., Pratt, A., Chen, R. S., et al. (1996). Cognitive profiles of difficult to remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology, 88*, 601–638.

Verhelst, N. D., & Glas, C. A. W. (1994). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models, foundations, recent developments and applications* (pp. 215–238). New York: Springer-Verlag.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *OPLM: Computer program and manual*. Arnhem: Cito.

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin, 101*, 192–212.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1994). Development of reading-related phonological processing abilities: New evidence of bidirectional causality from a latent variable longitudinal study. *Developmental Psychology, 30*, 73–87.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., Hecht, S. A., Barker, T. A., Burgess, S. R., et al. (1997). Changing relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: A 5-year longitudinal study. *Developmental Psychology, 33*, 468–479.

Yopp, H. K. (1988). The validity and reliability of phonemic awareness tests. *Reading Research Quarterly, 23*, 159–177.

Ziegler, J. C., & Goswami, U. C. (2005). Reading acquisition, developmental dyslexia and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin, 131*, 3–29.