Traditional and Computer-Based Screening and Diagnosis of Reading Disabilities in Greek

Athanassios Protopapas and Christos Skaloumbakas

Abstract

In this study, we examined the characteristics of reading disability (RD) in the seventh grade of the Greek educational system and the corresponding diagnostic practice. We presented a clinically administered assessment battery, composed of typically employed tasks, and a fully automated, computer-based assessment battery that evaluates some of the same constructs. In all, 261 children ages 12 to 14 were tested. The results of the traditional assessment indicated that RD concerns primarily slow reading and secondarily poor reading and spelling accuracy. This pattern was matched in the domains most attended to in expert student evaluation. Automatic (computer-based) screening for RD in the target age range matched expert judgment in validity and reliability in the absence of a full clinical evaluation. It is proposed that the educational needs of the middle and high school population in Greece will be best served by concentrating on reading and spelling performance—particularly fluency—employing widespread computer-based screening to partially make up for expert-personnel shortage.

Individual differences in response to school-based instruction are increasingly recognized as significant factors in the design and delivery of educational services. Educational policy dictates that students falling behind in academic achievement should be detected and properly supported with appropriate interventions. However, in Greece, there is no school-based system in general education for delivering assessment services to children in need of special learning accommodations. State-accredited service providers (usually child mental health agencies or Centers for Diagnosis, Assessment, and Support) diagnose students with learning disabilities (LD). Because child mental health providers are trained to deal with psychiatric disorders and not with academic failure, little attention has been generally paid to the educationally pervasive issue of LD. This state of affairs is partly responsible for the paucity of validated tools for diagnosis and treatment of learning problems, particularly in secondary education. Thus, most diagnosticians use qualitative means of assessment based on their personal experience.

The recent increase in societal awareness of LD has led to an increasing demand for diagnostic and remedial services and, hence, to a large population of students seeking assessment and receiving certification of LD status (usually referred to as a "dyslexia certificate"). A high proportion of schoolchildren referred for learning assessment are diagnosed with dyslexia, despite the lack of well-defined diagnostic criteria and of standardized tests and procedures. In the present study, we begin to address part of this gap by providing data on the types of skills in which children from the general school population differ from children referred for assessment (who will likely receive an LD diagnosis).

Specific Reading Disability

The term *specific learning disability* generally refers to academic failure in a specific domain, within the context of a structured educational system, in the absence of primary sensory, emotional, or behavioral disorders (Beitchman & Young, 1997). In a review of the evidence supporting the constructs related to various learning difficulties, Lyon, Fletcher, and Barnes (2002) listed six distinct subgroups of LD for which an adequate empirical basis has been gathered. Three of these subgroups are related to reading, including word recognition, comprehension, and fluency. The first of these, which concerns reading at the word level, is the most frequent, accounting for up to 90% of all LD cases. This is the subtype most appropriately termed *dyslexia*—a term often used synonymously with *reading disability* (RD) and the one on which we focus in this study, because of its preponderance.

Dyslexia is associated with difficulties in phonological processing (Wagner, Torgesen, & Rashotte, 1994; Wagner et al., 1997), typically assessed on a metalinguistic level (phonological awareness; McBride-Chang, 1995) and

as phonological short-term memory (Wagner & Torgesen, 1987). The deficit in reading performance is commonly considered to be the expression of an underlying deficit in the processing of the sound units that make up language (Shaywitz & Shaywitz, 2003; Wagner & Torgesen, 1987). Behaviorally, the dyslexic profile is typically associated with deficits in reading, spelling, and phonological processing (Warnke, 1999; Beitchman & Young, 1997; Shaywitz, 2003).

As most of the research on reading and RD has been conducted in English, a concern has arisen that findings may not be applicable to other languages—especially to languages with more consistent mappings between orthography and phonology—because the extreme inconsistencies found in English orthographies have been associated with very substantial differences from other languages in reading development and early reading performance (Landerl, Wimmer, & Frith, 1997; Seymour, Aro, & Erskine, 2003). Even moderate differences in grapho-phonemic regularity, as between Spanish and Portuguese, have been associated with differences in learning to read (Defior, Martos, & Cary, 2002). However, the importance of phonological processes in reading development—and their deficiencies in RD at various ages—have now been confirmed, by and large, in a number of languages with greater orthographic regularity than English, such as Czech (Caravolas & Volín, 2001), German (Mayringer & Wimmer, 2000), and Spanish (Jiménez González & Hernández Valle, 2000), among others. Thus, a cross-linguistic consensus is emerging that RD is an expression of a single dimension (Anthony & Lonigan, 2004; Schatschneider, Carlson, Francis, Foorman, & Fletcher, 2002) of poor phonological skills (Ramus, 2001; Ramus et al., 2003) across languages (Ziegler, Perry, Ma-Wyatt, Ladner, & Schulte-Körne, 2003).

An important issue that has emerged in the cross-linguistic conceptualization of dyslexia is the role of processing speed in the development and maintenance of reading automaticity. In languages with consistent orthography, such as Finnish (Holopainen, Ahonen, & Lyytinen, 2001), German (Landerl, 2001), Greek (Porpodas, 1999), Italian (Tressoldi, Stella, & Faggella, 2001; Zoccolotti, de Luca, Judica, Orlandi, & Spinelli, 1999), and Spanish (Jiménez González & Hernández Valle, 2000), it is the speed rather than the accuracy of reading (and of phonological component processes, including nonword decoding) that seems best to capture the essence of reading impairment. Reading fluency, defined as the speed of accurate oral reading of text, is taken to represent the automaticity of sublexical processes (van der Leij & van Daal, 1999; Wolf, Bowers, & Biddle, 2000) and, as such, is a crucial component of successful reading performance—in fact, an excellent index of overall reading competence (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Sabatini, 2002). Remediation studies have indicated that improvements in phonological awareness and decoding accuracy do not generally transfer to fluent reading of novel materials (Torgesen et al., 2001; Wolf & Katzir-Cohen, 2001; cf. Tressoldi et al., 2001). Thus, it remains an important question whether speed is a distinct dimension of performance related to reading and RD, and whether it is a dimension more critical to reading assessment than accuracy of reading or of tasks tapping component phonological skills.

Reading and Spelling Greek

Greek is considered relatively transparent (shallow) orthographically, occupying the second position in the classification of Seymour et al. (2003), after Finnish. There is, however, an asymmetry in orthographic transparency between spelling and reading: It is possible to *read* the majority of words correctly based on straightforward decoding, but it is impossible to *spell* correctly based on the words' pronunciation alone.

In particular, the phonology-to-orthography conversion is not at all predictable for three of the five vowels (Mavrommati & Miles, 2002), because they can be spelled in several ways, with one or two letters (e.g., /e/ is spelled as ε or αι). The correct spelling depends on the morphology and etymology of the word (Chliounaki & Bryant, 2002; Porpodas, 1999). Consonant spelling is regular, but not always simple. For example, voiced stops and all palatals are spelled with combinations of two or more letters (e.g., /ç/ can be spelled χι, χει, χοι, etc.).

In contrast to spelling, reading is very regular, because the orthography-to-phonology conversion is systematic, if not entirely simple. There are many letter combinations with special phonetic value (e.g., /u/ is spelled ου) and special conditions (e.g., υ is pronounced /v/ when it follows an unstressed α or ε), but they are regular, and their combined effect on pronunciation is predictable. One significant exception to reading regularity concerns consonant palatalization. Specifically, when one of the six /i/ graphemes follows a consonant and precedes a vowel, it may indicate (a) palatalization of the preceding consonant (μάγια→/ˈmaʝa/); (b) pronunciation of /i/ (αίτια→/ˈetia/); (c) both of the above (άγια→/ˈaʝia/); or (d) pronunciation of a palatal consonant (σπίτια→/ˈspitça/). The correct reading is lexically determined.

In sum, Greek spelling is irregular and complex, primarily for vowels, owing to morphological categories and historical origins of the words. Lexical and morphological knowledge is thus necessary for correct spelling. In contrast, Greek reading is mostly regular and, with a few exceptions, can be mastered entirely on the basis of grapho-phonemic conversion rules (though some of the rules can be complicated). Therefore, if the RD profile were primarily a reflection of language-specific mapping consistency, Greek children with RD would be expected to have primarily problems with spelling accuracy, whereas their reading skills

should present little difficulties. However, taking into account the findings from other languages with consistent orthography, and without discounting the importance of phonological awareness and accurate decoding, it is the *speed* of processing that should emerge as a critical dimension for assessment, and this would be expected to affect reading performance despite the high graphophonemic regularity.

Assessment Measures

As a diagnostic starting point, isolated word reading and text reading accuracy and speed must be included in any testing battery as "surface" indices of RD (Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003). Inaccurate reading is a defining sign of RD, particularly in orthographically opaque languages such as English (Landerl et al., 1997; Ziegler et al., 2003), but may lose discriminating power with age (Bruck, 1992; Ellis et al., 2004). In contrast, as previously mentioned, speed measures seem to adequately capture the magnitude of difficulties, particularly in orthographically transparent languages, but also increasingly in English (see Fuchs et al., 2001; Wolf & Bowers, 1999).

Spelling ability is rarely spared in the presence of reading problems (Porpodas, 1999; Wimmer & Mayringer, 2002). In Greek, most spelling problems partly stem from inflection rule violations (resulting in "morphological" errors) and partly from deficits in awareness of the etymology of the word and the ability to relate it semantically to already known words (resulting in "historical" errors). Phonologically inaccurate spellings are rare past the first or second grade, though not entirely absent even in orthographically transparent languages, such as Czech (see Caravolas & Volín, 2001). Text spelling can tap syntactical and semantic information for disambiguation, whereas for isolated words one must rely on a combination of lexical and derivational knowledge.

With respect to phonological awareness, we chose phoneme deletion because it is one of the most difficult tasks, in an effort to uncover possible discriminating variables at an age where phonological processing difficulties may have subsided (see McBride-Chang, 1995). The importance of assessing phonological awareness past the age of primary education is unclear (Shaywitz, 2003). In some languages with more consistent orthography, such as German (Landerl & Wimmer, 2000) and Italian (Cossu, Shankweiler, Liberman, Katz, & Tola, 1988; Zoccolotti et al., 1999), most children with RD display phonological problems early on, but then seem to attain high absolute levels of phonological awareness even though their reading performance remains significantly below that of good readers (the situation for Finnish remains unclear; see Holopainen et al., 2001; Muller & Brady, 2001).

Pseudoword reading has been one of the most widely used indicators of RD (Rack, Snowling, & Olson, 1992), as it samples a child's ability to phonologically decode a word without the help of orthographic knowledge, both in English (Siegel, 1989; Stanovich, 1988) and in languages with more consistent orthographies (Lehtola & Lehto, 2000; Jiménez González & Hernández Valle, 2000; Tressoldi et al., 2001).

To examine phonological memory, we used digit span, a frequently employed marker of LD (Torgesen & Houck, 1980; Wagner et al., 1997), and pseudoword repetition, a marker of language-based disabilities (Gathercole, Willis, Baddeley, & Emslie, 1994; Bishop, North, & Donlan, 1996). Besides digit span, general intelligence measures continue to figure prominently in testing practice, despite longstanding criticisms of their validity in LD diagnosis (Fletcher et al., 2002; Lyon et al., 2002; Siegel, 2003). Hence, in the particular case of Greek, it is important to examine the potential role of such measures in the context of an extensive assessment battery.

Auditory speech discrimination ability, often found to be deficient in children and adults with RD (Adlard & Hazan, 1998; De Weirdt, 1988; Gottardo, Siegel, & Stanovich, 1997), was assessed in a task requiring same–different discrimination of pseudoword pairs (see Schulte-Körne, Deimel, Bartling, & Remschmidt, 1999; Serniclaes, Sprenger-Charolles, Carre, & Demonet, 2001). Finally, in the computer-based assessment, it was also possible to include *nonverbal* auditory perception measures (which require precise adaptive control of stimulus parameters and, thus, cannot be administered by a human), because such skills have been associated with language development and with language-based LD in children and adults (Ahissar, Protopapas, Reid, & Merzenich, 2000; De Weirdt, 1988; Reed, 1980; Tallal, 1976, 1980).

Goals of Assessment

One important distinction must be made between the goals of a "traditional" clinical assessment, administered with paper and pencil in a one-on-one interview, and those of a computer-based assessment, administered and scored automatically without the need for human supervision or intervention. Clinical assessment tasks are administered and scored by expert personnel in the course of a comprehensive evaluation aiming to determine the appropriate diagnosis or educational intervention for the child being tested. Such comprehensive evaluation will typically not be restricted to cognitive and achievement measures, but may encompass additional parameters, such as motivation (Sideridis, Morgan, Botsas, Padeliadu, & Fuchs, 2006). The human expert (typically an educational psychologist or special education teacher) uses the quantitative information provided by the assessment scales, in conjunction with other observations, measures, and history, and arrives at a clinical diagnostic judgment. The aim of a "traditional" assessment battery in this context, then, is to provide the special-

ist with reliable and valid diagnostic information.

In contrast, automated computer-based assessment cannot lead to an automatic diagnosis or similar judgment, because it lacks many elements that are essential to the diagnosis (history, emotional and behavioral evaluation, etc.), and also because of the lack of total control over the process of data collection. As a simplistic example, it is possible that a child is very distracted when carrying out the computer-based tasks, leading to poor performance. This cannot be ascertained with any confidence in the automated procedure, but would hardly escape the notice of a competent clinician. Therefore, the aim of a computer-based assessment can be to serve only as a type of *screening* procedure, on the basis of which children with certain performance profiles (typically, but not necessarily, poor performance on critical tasks) are referred for comprehensive evaluation by specialist professionals.

Thus, although the word *assessment* is used here to refer both to the battery of clinically administered tests and to the computer-based tests, there should be no confusion about the sharp distinction in the role and purpose of each type of assessment (i.e., diagnosis vs. screening). Our purpose is not to compare the two types of assessment, but rather to examine their usefulness in their separate roles using a common approach.

Study Objectives

In the study reported here, we attempt to answer the following questions:

1. *What are the characteristics of the RD population in Greece?* A very large proportion of children seeking assessment due to difficulties at school end up with a diagnosis of RD. Therefore, the comparison of the general population with a self-selected clinical sample seeking services can provide important clues to the main features of RD in Greece. This method offers a non-circular and unbiased perspective on what the features of RD are, given the current practice and educational situation.
2. *What are the main relevant dimensions on which children actually differ?* Given a battery of largely intercorrelated assessment tests, it is important to determine the independent contribution of distinct types of skills to overall profiles. This has both diagnostic and theoretical consequences, as it restricts and directs the number and type of constructs that need to be assessed, as well as the type of tasks to assess them with.
3. *What is the empirical basis for expert judgment on RD diagnosis?* That is, what aspects of a comprehensive assessment do professionals most attend to when deciding to diagnose a child with RD? Given a comprehensive assessment battery and two or more independent judgments regarding RD diagnosis from experienced professionals, it is possible to identify the patterns in assessment data that are most strongly associated with the expert judgment.
4. *Is computer-based screening psychometrically adequate?* Once the relevant dimensions and tasks are known, it is possible to devise automated data collection procedures for skill assessment, aiming to provide first-pass screening services to the school population. Strict psychometric evaluation of the screening procedure is a necessary (but not sufficient) condition before its widespread adoption can be recommended.

These questions are addressed with two sets of measures on a medium-size school sample from the seventh grade. Because of the diversity of research questions, age was kept constant to reduce variance not associated with the variables of interest. The age of 12 to 13 years was chosen for educational and practical reasons. Specifically, the Greek educational system provides for special treatment (i.e., oral instead of written final examinations) no earlier than seventh grade, which is also the time of transition from a more lenient and less competitive to a more impersonal and demanding environment, which accentuates already existing literacy-related difficulties. Therefore, the majority of children seeking LD assessment are from the seventh grade, when the outcome of the evaluation becomes highly relevant, resulting in an increased need for empirically supported assessment at this age.

Method

Participants

In total, 261 seventh-grade children contributed data to the analyses reported here, after removing cases with missing data points and taking into account additional considerations, as described in detail in the data preparation section. Of these, 185 were recruited in eight public schools, selected to cover a wide range of socioeconomic status (SES), in the province of Attiki (which includes the general Athens metropolitan area) and one other large city (Patra). This first subgroup (henceforth *school sample*) constituted our general population sample, and completed all assessment tasks. The second subgroup, consisting of 28 children (*clinical sample*) were recruited at the Medical-Pedagogical Center of the Children's Psychiatric Hospital, to which they were referred for the assessment of learning difficulties as a result of poor academic performance. This group completed all assessment tasks after being examined by a child psychiatrist and a psychologist to rule out other medical or psychiatric conditions, primary behavioral or emotional problems, or low intelligence. The third subgroup (*retest sample*), including 48 children, were recruited at a private

school, to participate only in the test–retest reliability assessment of the software, and completed only the computer-based tests. These children did not provide data for any of the other analyses.

Table 1 shows, for each group, the number of children of each gender, their age, and an estimate of nonverbal intelligence (for the school and clinical samples). There was no difference between the school sample and the clinical sample in either age or nonverbal intelligence, both $F(1, 211) < 1$; there was also no significant difference in gender proportions between these two groups, $\chi^2(1, N = 213) = 3.0$, exact $p > .1$, two-tailed.

Testing Materials

Traditional Assessment. Traditional assessment included some preexisting tests and some new tests with materials constructed for this purpose. The preparation of the test battery aimed to achieve the effects of a typical assessment, as practiced in Greece, with up-to-date, properly controlled and justified content where no standardized materials exist. The complete battery was named KLIMA, from the Greek words for "learning assessment scale" (κλίμακα μαθησιακής αξιολόγησης). It included the following measures:

Pseudoword reading. Twenty pseudowords of two to five syllables long (Maridaki-Kassotaki, 1998) were presented on a sheet of paper for the child to read aloud. The total reading time and number of incorrectly read items were noted.

Pseudoword repetition. The remaining 20 pseudowords from Maridaki-Kassotaki (1998) that were not used in the reading task were pronounced by the experimenter one by one for the child to repeat. The number of incorrectly repeated items was noted.

Word reading. Children read aloud a list of 84 words, one to seven syllables in length, presented in three columns on a sheet of paper. The total reading time and number of incorrectly read items were noted. Words were chosen according to the following criteria:

1. Grammatical category. Most items were content words, and the ratio of verbs to nouns was about 1:3.
2. Length. 44% three-syllable words, 21% two-syllable, 15% each one- and four-syllable, and a few longer items were included.
3. Printed frequency, based on the Hellenic National Corpus (HNC; Hatzigeorgiu et al., 2000). Medium-frequency items, with 12 to 99 appearances in 32 million, made up the majority, whereas high- and low-frequency items were also included.

Care was taken to include a variety of representative declension and conjugation types expressed in the words' suffixes, because Greek relies on inflections to carry meaning, affecting readers' allocation of attentional resources (Chitiri & Willows, 1994).

Text reading and comprehension. Three passages of increasing difficulty, 72, 90, and 78 words long, were read aloud by the students, and their total reading time and total number of reading errors were noted. Following each passage, comprehension questions were asked, and points were given for correct responses (partial points were given for approximations). For the second and third passage, 1-minute silent study was provided after reading aloud and before the questions.

Passages were composed according to the following criteria:

1. Genre. By the Stein (1984) classification, the three passages were of the reaction-sequence type I, goal-directed with obstacle, and goal-directed with no obstacle, respectively.
2. Vocabulary. Longer and less frequent words were used increasingly from the first to the third passage.
3. Syntactic complexity. Longer sentences and subordination were used increasingly in the second and third passages.

The comprehension questions probed (a) information memorization, putting more weight on essential pieces of information; (b) information integration from disparate points in the passage; and (c) reasoning on the basis of information (Hannon & Daneman, 2001).

Word spelling. A list of 21 words was dictated at a child-determined rate, and the total number of spelling errors was noted (more than one error per word was possible). Words were chosen to be frequent and to provide opportunities for a variety of spelling errors, primarily morphological (i.e., on the inflectional suffix) but also his-

TABLE 1
Age, Gender, and Nonverbal Intelligence by Student Group, Including All Children with Full Data Sets

	Gender		Age (months)		RSPM (raw)	
Sample	Boys	Girls	*M*	*SD*	*M*	*SD*
School	93	92	150.5	5.1	37.1	9.7
Clinical	19	9	149.6	4.0	37.4	7.4
Retest	26	22	152.3	3.8	—	—

Note. $N = 261$, after exclusion of participants with invalid or missing data. RSPM = *Raven's Standard Progressive Matrices* (Raven, 1976).

torical (i.e., on the root, related to the word's etymology).

Text spelling. A 49-word passage from Zahos and Zahos (1998) was dictated at a child-determined rate, and the total number of spelling errors was noted. The passage contained well-known words, and its meaning was easy for the target age; however, the words provided many opportunities for errors, primarily morphological but also historical.

Raven's Standard Progressive Matrices. The full 60-item version of the *Raven's Standard Progressive Matrices* test (RSPM; Raven, 1976) was used. The number of correct choices (raw score) and the time to complete the test were noted.

Digit span. The Digit Span subscale from the Greek standardized version of the *Wechsler Intelligence Scale for Children*, third edition (WISC-III; Georgas, Paraskevopoulos, Bezevegis, & Giannitsas, 1997), was used, including forward and backward span, following standard administration procedure and termination criterion. The total number of sequences reproduced correctly (raw score) was noted.

Arithmetic. The Arithmetic subscale from the Greek standardized version of the WISC-III (Georgas et al., 1997) was used, following the standard instructions (for the target age, administration begins with Item 6) and termination criterion. The number of correct responses (plus 5) was noted (raw score).

Phoneme deletion. A set of 22 two-syllable and three-syllable pseudowords were constructed following Greek phonotactic structure, including a high proportion of consonant clusters. For each pseudoword, one phoneme was the designated deletion target, varying greatly in phoneme type, word position, and syllabic position. Each pseudoword was presented orally and, once repeated correctly, was presented again along with the phoneme to be deleted. The total number of incorrect responses was noted.

Speech sound discrimination. This is a subscale from the AthenaTest for LD (Paraskevopoulos, Kalantzi-Azizi, & Giannitsas, 1999). It includes 32 pseudoword pairs, 8 of which are identical repetitions. The remaining 24 item pairs differ by a single phoneme modification, insertion, or translation. Each pair was presented orally to the student to be judged as *same* or *different*. The total number of incorrect responses was noted.

Computer-Based Assessment. The computer-based assessment was designed to achieve automated screening—that is, referral of schoolchildren for full assessment by expert personnel. Thus, it included a number of tasks known to correlate with the presence of LD, particularly with specific reading disabilities.

Text reading. Ten passages were constructed, 30 to 61 words long, of increasing difficulty owing to sentence length, subordination, passive voice, and connections between individual passages. The time to read each passage was noted, based on the mouse clicks that initiated and terminated its presentation. Comprehension was assessed by presenting, after each passage, four images (line drawings). One of the drawings depicted the situation described in the passage, whereas the other three contained increasingly less important inaccuracies. Thus, the overall difficulty resulted from the combination of passage complexity and salience of information represented in the image differences. The total comprehension score was the number of correct image selections.

Text spell-checking. Nine passages, 28 to 46 words long, were constructed with a total number of 27 spelling errors (14 morphological and 13 historical), distributed heterogeneously (1–9 errors per passage). Each passage was presented on the screen for the child to detect the errors (by clicking on them) and correct them (by clicking on the appropriate letters in an on-screen keyboard display). The total number of errors in the resulting passages was noted, as well as the total time spent on the task.

Frequency discrimination. An adaptive psychophysical procedure based on Treutwein (1995) was used to determine the minimum frequency difference needed to judge that two 250-ms long pure tones 500 ms apart were different. Tone frequencies were centered on 1 kHz and differed by up to 600 Hz. Stimulus parameters were based on Ahissar et al. (2000).

Tone sequencing. The same adaptive psychophysical procedure was used to determine the minimum temporal separation between two successively presented 20-ms pure tones needed to correctly reconstruct their sequence. The frequency of each tone was either 800 or 1200 Hz, and the interval between them started at 500 ms. In a second phase, three-tone sequences were used. Stimulus parameters were based on Ahissar et al. (2000), and the general design can be traced back to the original Tallal repetition test (Tallal, 1976).

Pseudoword dictation. A total of 23 pseudoword stimuli were made up of 6 CV (consonant-vowel) syllables, 6 CVCV bisyllables, 3 CVCVCV three-syllable items, and 4 two-syllable and 4 three-syllable complex pseudowords containing one to three consonant clusters. The 15 simple items were constructed using only /p/, /t/, /k/, and /a/, whereas a great variety of phonemes was used for the 8 complex items. Each item was presented auditorily, and the appropriate response was a phonologically correct spelling (in some cases, more than one was possible). The number of correct responses was noted.

Word–picture matching. Thirty sets of 4 items each were constructed, one of the 4 items in each set being a correctly spelled word, for which a corresponding line drawing was made. The other 3 items of each set were either phonologically similar words or phonologically identical misspellings of the correct word. In each trial, the 4 items and the corresponding sketch were presented visually for the child to select the correct word. The number of correct responses was noted.

Letter span. This task was modeled after the standard digit span task, but used letters instead of numbers. A random set of Greek uppercase consonants made up the sequence of the desired length for each trial, presented visually at the rate of one letter per second. The child had to reproduce the correct sequence by clicking on the letters with the mouse. Two sequences of each length were presented, starting at length 2 and increasing by 1 if at least one of the two was responded to correctly. The total number of correctly reproduced sequences was noted.

A cartoon character, "Professor Vidas," bearing a remote resemblance to Albert Einstein, provided the situations in which the child was asked to offer help (with labeling pictures, correcting writings, catching mice, mating birds, etc.). Before each task, instructions were spoken through the headphones while at the same time the intended actions were demonstrated visually on the screen. Instructions were recorded by a professional at an age-appropriate tone. During each task, the child was only given options compatible with the intended action, activating buttons and screen regions only when relevant. Guiding visual cues were provided to attract attention to the next expected action (e.g., to initiate the next trial). There was no feedback on user performance except for the adaptive psychophysical tasks, in which feedback was necessary for stable convergence to the threshold.

The software was named *eMaDys*, from the Greek words for "detection of learning difficulties" (εντοπισμός μαθησιακών δυσκολιών). The design of the software was based on the principles of simplicity and automation. *Simplicity* refers to the presence of only functional elements in the tasks. Both graphic design and interaction minimized distraction, allowing the user to focus on the task. Navigation was strictly linear, without any options or choices to be made other than for responding to the assessment items. All user input was based on the computer mouse. When typing letters was necessary, a virtual keyboard appeared on the screen, and the letters were selected using the mouse.

Automation means that no human intervention is necessary at any stage of the assessment, either to administer or to score the tests. Software installation requires no selections of the user and is completed in a short series of clicking on *Next*. Test administration requires no supervision: A child sits in front of the computer, puts on the headphones, enters a numerical ID given by the teacher (so that no name or other personal information is associated with the results file), and follows the spoken instructions through the tasks, which advance automatically once completed. Responses are scored by the software, and referrals can be generated automatically as soon as the testing process is completed. The requirement for automation precludes spoken or handwritten responses, which would necessitate the presence of a human rater. Because such presence is not guaranteed, and, in fact, a computer-based screening solution is most needed where such presence is *not* available, full automation was considered critical for the viability of the project. Thus, many potentially useful tasks, such as standard phonological awareness tasks, could not be included in the screening software.

The combined result of simplicity and automation is *ease of use*: Any teacher can install the software, and any child (of the intended age) can complete the assessment tasks, regardless of any prior knowledge related to computers.

Procedure

Each child was assessed on the two batteries (traditional and computer-based) on separate days. The order of administration of the two batteries was determined primarily by the need to accommodate scheduling without disrupting school activities (e.g., availability of computer laboratory) while keeping the total time to a minimum. Typically, some children would complete the computer-based battery first, while others completed the traditional battery; the next day each child would complete the other battery. Thus, there was no overall systematic order, although no attempt was made to counterbalance or randomize the order of administration. Administration of each battery took 60 to 80 min, spanning two class periods, depending on individual speed and performance. Traditional assessment was tape-recorded, except for Raven's SPM and the two spelling tasks.

For the school sample, the traditional assessment was administered by a trained university student (senior or graduate) in a quiet room at the school. A break was offered halfway through testing or if the child became obviously fatigued or restless (although this was uncommon). Traditional assessment was always administered individually. Computer-based assessment took place at the school computer laboratory, at times when no class was scheduled in it, administered simultaneously to as many children as could use the available computers in the laboratory (usually 4–8; up to 12 in one school). Care was taken to prevent viewing of adjacent screens when multiple children used the screening software at the same time. Children wore closed-type headphones while using the software. Breaks were offered between tasks but were rarely taken by the children, who preferred to continue "playing" at the computer. For the children taking part in the reliability study, computer-based assessment was administered twice, in exactly the same way, with 3 to 5 weeks between the two administrations. For the clinical sample, all testing took place in the tester's office at the hospital, as part of the children's comprehensive assessment procedure.

Results

Data Preparation

The following procedure was performed on the data set from the two

main participant groups, including the school sample and the clinical sample. Of the 270 children originally recruited, age was not available for 9 children and was above 168 months for 9 more children (likely repeating the class). Nine children spoke a native language other than Greek (immigrants), 2 children (from the clinical sample) were identified as having marginal intelligence, and 1 child was missing several measures (incomplete administration) on both batteries. These 30 children were removed from the sample.

Traditional (15 variables) and computer-based assessment data (10 variables) were then temporarily separated. Individual missing value analysis indicated 0.64% missing values for the traditional assessment (23 data points total, proportions ranging between 0.0% and 2.1% per variable) and 0.37% missing values for the software (8 data points total, up to 1.4% per variable). These missing values were replaced using the Expectation Maximization method, and then the two datasets were recombined.

Due to scheduling difficulties, 23 children had not completed both batteries: 22 children with no data from the computer-based assessment, and 1 child lacking traditional assessment data were removed from further analysis, leaving 217 participants.

Inspection of descriptive statistics, histograms, and Q-Q plots indicated significant to severe deviations from normality in most measures. Variables were transformed with the usual functions (inverse, log, and square root, with linear shifting when necessary to avoid negative arguments) to bring their skewness and kurtosis statistics within one corresponding standard error, taking into account the visual inspection of the histogram in selecting transformations. All variables were thus transformed, and all subsequent analyses refer to the transformed variables.

Four multivariate outliers were identified based on the Mahalanobis distance statistic and were verified on scatterplots. The corresponding 4 participants were removed from the sample. No combination of variables met the multicollinearity criteria for rejection; therefore, no further modifications to the data set were made. Thus, the final combined school and clinical samples, as shown in Table 1, included 213 children in total, providing data points to 25 variables.

Regarding the retest sample, out of 51 children originally participating, 3 participants were removed: 2 for missing the second administration due to scheduling difficulties, and 1 for missing three data points due to software failure. Thus, 48 children formed the final retest sample used in the analysis, as shown in Table 1. All data from these children were transformed using the same functions used for the school and clinical sample, without further modification.

In a preliminary MANCOVA of the 25 measures, including data from the 213 children in the school and clinical samples, with sample and gender as two-level factors and age as a covariate, there was no significant effect of age (as expected, due to the restricted age range) and no significant effect of gender except for the spelling measures, with boys making (or accepting) more mistakes than girls: word spelling errors, $F(1, 208) = 5.77$, $p = .017$; text spell-checking, $F(1, 208) = 6.49$, $p = .012$; word–picture matching, $F(1, 208) = 5.94$, $p = .016$; none of which interacted significantly with sample ($p > .3$; see Note 1). Therefore, age and gender were not taken into account in any of the following analyses.

Comparison of School and Clinical Samples

To determine the measures that best distinguish children with RD from the general population, we should ideally compare a pure RD sample to a non-RD sample, so that the corresponding means and variances would approach true estimates of the RD and non-RD population statistics (contingent on sampling adequacy). The effect sizes would then constitute good estimates of the discriminant usefulness of each measure. In our case, the school sample included a small number of children with a profile of RD, and the clinical sample included a few children who were not diagnosed with RD after comprehensive assessment. Therefore, the comparison of these two groups does not correspond to a perfect separation of samples with and without RD. However, as stated previously, there is no standardized tool to diagnose RD independently of the measures administered in this study. Identification of RD children hinged, in part, on the measures described earlier. Thus, it would be circular to compare children with and without RD on the same measures. On the other hand, the expected proportions of children with RD in the two samples are very different: Experience suggests that up to 10% of the school sample and perhaps 80% or more of the clinical sample have LD. Therefore, by comparing these two samples on the various measures, we can statistically identify the measures that best separate the intended populations of children with and without RD. The effect sizes will approximately indicate the ranking of measures in identifying children with RD, even though they will not constitute accurate estimates of the hypothesized pure group differences.

Table 2 shows the results of one-way ANOVAs between the school and clinical sample for each transformed variable. Because many differences are highly statistically significant, the effect size (partial η^2 and Cohen's d) is reported as a more appropriate index of the relevance of each measure to the identification of RD (Ives, 2003; Onwuegbuzie, Levin, & Leech, 2003). In the traditional assessment, it is clear that the school sample and the clinical sample differed primarily in reading speed, secondarily in reading and spelling accuracy, and little if at all in more general cognitive measures or in phonetic and phonological measures. Similar computer-based measures could also distinguish the two samples, pri-

TABLE 2
Percentile Values of Untransformed Variables for All Traditional and Computer-Based Assessment Measures and One-Way ANOVA of the Corresponding Differences in Transformed Variables

Variable	School sample[a]			Clinical sample[b]					Effect size	
	10	50	90	10	50	90	*F*(1, 211)	*p*	Partial η^2	Cohen's *d*
Traditional Assessment										
Pseudoword reading errors	1	4	10	4	7	13	20.46	< .0005	.09	.92
Pseudoword reading time(s)	28	41	64	44	59	128	37.69	< .0005	.15	1.25
Pseudoword repetition errors	1	4	8	1	4	9	.01	.922	.00	.02
Word reading errors	0	1	8	2	7	21	44.63	< .0005	.17	1.36
Word reading time(s)	68	89	132	101	132	212	42.33	< .0005	.17	1.32
Text reading errors	0	4	12	1	9	16	13.56	< .0005	.06	.75
Text reading time(s)	83	102	151	138	171	277	83.49	< .0005	.28	1.85
Text comprehension score	7	10	16	6	9	12	7.09	.008	.03	.54
Word spelling errors	0	1	6	1	6	12	26.67	< .0005	.11	1.05
Text spelling errors	0	3	15	4	13	26	32.25	< .0005	.13	1.15
RSPM (raw score)	23	39	48	27	38	48	.01	.940	.00	.02
WISC-III Digit Span (raw score)	10	13	18	9	12	15	5.28	.023	.02	.47
WISC-III Arithmetic (raw score)	14	17	21	14	16	19	4.08	.045	.02	.41
Phoneme deletion errors	2	5	11	2	7	13	3.95	.048	.02	.40
Speech discrimination errors	2	5	10	2	5	11	.36	.548	.00	.12
Computer-Based Assessment										
Text comprehension score	5	8	9	6	7	9	1.73	.190	.01	.27
Text reading time(s)	11	19	33	19	26	50	27.05	< .0005	.11	1.05
Text spell-checking errors	8	17	28	17	27	33	30.18	< .0005	.13	1.12
Text spell-checking time(s)	43	61	91	49	79	149	20.06	< .0005	.09	.91
Frequency discrimination (Hz)	28	62	213	31	50	274	.04	.841	.00	.04
Two-tone sequencing (ms)	0	78	451	0	91	405	.02	.875	.00	.03
Three-tone sequencing (ms)	51	176	563	56	207	447	.02	.891	.00	.03
Pseudoword dictation score	16	20	22	13	18	21	5.91	.016	.03	.49
Word–picture matching score	20	25	28	13	20	25	28.20	< .0005	.12	1.08
Letter span score	4	6	8	3	4	7	12.62	< .0005	.06	.72

Note. *N* = 213. Untransformed variables derived from the 15 clinical assessment measures and 10 computer-based assessment measures are expressed in percentile values at the 10th, median (50th), and 90th percentile. Effect sizes are absolute values. RSPM = *Raven's Standard Progressive Matrices* (Raven, 1976); WISC-III = *Wechsler Intelligence Scale for Children*, 3rd ed., Greek version (Georgas et al., 1997).
[a]*n* = 185. [b]*n* = 28.

marily the two spelling tasks, followed by reading speed.

Structure of Assessment Batteries

The fact that our RD-rich clinical sample differs from the general population on a set of measures does not necessarily indicate clearly that there are commonalities among the types of skills underlying RD status, because of the high correlations among the measures. Therefore, exploratory factor analyses were performed to examine the underlying structure of the assessments, in order to (a) identify potential latent variables indexing relevant abilities, and (b) determine the extent to which traditional and computer-based assessments measure similar constructs. Because no relevant empirical basis exists for Greek—especially for children of this age—on which to construct specific hypotheses to be tested (e.g., with confirmatory factor analyses), it was judged necessary to begin with exploratory analyses in this study.

The intercorrelations among the 15 traditional assessment variables are shown in Table A1 in the Appendix. For these variables, principal axis factoring with Varimax rotation revealed three factors with initial eigenvalues greater than 1.0, together accounting for 54% of the total variance. Extraction communalities ranged between .21 (for pseudoword repetition) and .86 (for word reading time), but only for RSPM and digit span raw scores did they fall below .4 (along with pseudoword reading time).

Inspection of the rotated factor matrix (see Table 3), with only loadings above .4 interpreted, suggests that the traditional assessment roughly addresses (a) reading and spelling accuracy; (b) reading fluency, a construct that refers to the speed of accurate reading (see Note 2); and (c) general mental ability ("intelligence"). It is no-

TABLE 3
Rotated Factor Loadings of the Transformed Variables from the Traditional Assessment for School and Clinical Samples

Variable	Factor 1	Factor 2	Factor 3
Word reading errors	**.70**	**–.40**	.19
Word spelling errors	**.66**	–.33	.32
Pseudoword reading errors	**.66**	–.25	.22
Text spelling errors	**.64**	**–.46**	.28
Text reading errors	**.61**	**–.40**	.16
Phoneme deletion errors	**.51**	–.17	**.40**
Word reading time	–.34	**.85**	–.16
Text reading time	–.37	**.81**	–.26
Pseudoword reading time	–.29	**.78**	—
Speech discrimination errors	.25	—	**.59**
WISC-III Arithmetic raw score	–.28	—	**–.59**
Text comprehension score	–.15	.16	**–.56**
RSPM raw score	.28	.14	**.52**
WISC-III digit span raw score	–.18	.28	**–.50**
Pseudoword repetition errors	—	—	**.45**
Score covariance	.75	.88	.71
Sum of squared loadings (% var)	20.1	19.0	15.2

Note. *N* = 213. Loadings 0.4 and higher are shown in bold. Values of less than 0.1 are not shown at all.

TABLE 4
Rotated Factor Loadings of the Transformed Variables from the Computer-Based Assessment for School and Clinical Samples

Variable	Factor 1	Factor 2	Factor 3
Three-tone sequencing	**.85**	–.14	—
Two-tone sequencing	**.83**	–.14	—
Frequency discrimination	**.55**	—	—
Word–picture matching score	—	**.77**	**.42**
Text spell-checking errors	–.11	**.65**	**.48**
Letter span score	–.30	**.49**	.25
Text comprehension score	—	**.43**	–.13
Pseudoword dictation score	–.33	.36	.15
Text reading time	—	.20	**.69**
Text spell-checking	—	—	**.56**
Score covariance	.85	.73	.63
Sum of squared loadings (% var)	19.5	16.5	13.0

Note. *N* = 213. Loadings 0.4 and higher are shown in bold. Values of less than 0.1 are not shown at all.

table that the two somewhat "auditory" tasks (i.e., speech discrimination and pseudoword repetition) group with the more clearly "intelligence" measures in the third factor.

For the 10 computer-based assessment variables, the analysis again revealed three factors with initial eigenvalues greater than 1.0, together accounting for 49% of the total variance. Extraction communalities ranged between .21 (for text reading comprehension) and .78 (for word–picture matching). Inspection of the rotated factor matrix (see Table 4) suggests that the computer-based assessment roughly addresses (a) auditory skills, (b) reading and spelling accuracy, and (c) text processing speed. It is notable that the two spell-checking accuracy tasks also group with processing speed in the third factor.

Table 5 shows the correlations between the regressed factor scores from the two analyses (traditional vs. computer-based). The highest correlations are observed among the factors thought to assess similar constructs, although the separation between accuracy and fluency is not perfect. Notably, the auditory factor from the computer-based assessment is significantly correlated with the intelligence factor from the traditional assessment. To confirm the relation between the factors derived from each set of measures, the 25 variables were analyzed together, resulting in five factors, together accounting for 53% of the total variance. A four-factor solution would leave several variables not loading on any factor, whereas a six-factor solution could not achieve convergence. The rotated factor matrix (not shown) had variables from the accuracy and speed factors of the two assessments grouped together, whereas the variables from the auditory and general cognitive factors remained separated.

In sum, it appears that both batteries assessed a similar pair of constructs related to reading and spelling fluency (primarily speed) and accuracy. However, the software battery also derived an auditory factor that

was not very strongly related to any of the traditional measures. Taking into account the comparison between school and clinical samples (see Table 2), this auditory factor is probably not relevant to the identification of children with RD. Similarly, the traditional battery derived a mental ability factor, which is also apparently not important for this type of classification. Perhaps the measures loading primarily on these factors are useful for different classifications or clinical evaluations.

TABLE 5
Correlations (*r*) Between Regressed Factor Scores for Traditional Assessment (Rows) and Computer-Based Assessment (Columns) for the Combined School and Clinical Sample

	Computer-based assessment		
Traditional assessment	**Factor 1 (Auditory)**	**Factor 2 (Accuracy)**	**Factor 3 (Text processing)**
Factor 1 (Accuracy)	−.119	**.565**	**.373**
Factor 2 (Fluency)	−.088	**−.281**	**−.614**
Factor 3 (Intelligence)	**−.345**	**.378**	.198

Note. $N = 213$. Correlations significant to $p < .0005$ are shown in bold.

Basis of Expert Judgment

How do skilled professionals evaluate measures such as those in the traditional assessment to reach a diagnosis of RD? Although all measures used are internationally considered standard components of an RD testing battery, it is not known (a) which variables are most relevant to diagnosis in the context of the Greek language and educational system or (b) to what extent all of the measures are actually taken together into consideration for the comprehensive evaluation.

Criterion for Discriminant Analysis. The following analysis was performed on a subsample including 134 children (64 girls) from the school sample only. No children from the clinical sample were included because much additional information was already known about them, and this would likely interfere with the experts' judgment. Two experienced professionals in the diagnosis and remediation of RD (including one of the authors) examined each individual assessment sheet, with the results of all the aforementioned traditional assessment tasks, and determined whether, on the basis of these data, the child should be referred for comprehensive RD assessment or not. None of the children had been administered the assessment tasks by these experts; thus, the only available basis for this judgment was the written record of the traditional assessment battery. The two experts worked independently.

One hundred five (105) children were identified by both experts as not likely to have reading difficulties (no disabilities; ND group), and for 9 children (2 girls), both experts indicated with confidence that they should be referred for full evaluation (RD group). The remaining 20 children (9 girls) were identified for referral by only one of the two experts. Thus, the overall interrater agreement among the two experts was 85% (concordance rate $\kappa = .40$; when relevant, concordance rates are henceforth reported using Cohen's κ statistic). The 20 ambiguously identified children were subsequently classified as either ND ($n = 6$) or RD ($n = 14$) at a second stage, after discussion and agreement among the two experts. The two classifications (with 9 unambiguous RD and 20 after discussion and reconsideration) were taken as the reference criteria for the following analyses.

Relevance of Traditional Assessment Measures for RD Classification. Taking into account only the unambiguous classification (105 + 9 children), stepwise discriminant analysis of the 15 traditional assessment measures (with group classification probabilities derived from the sample) resulted in a function correctly classifying 100% of the children (99.1% in "leave-one-out" cross-validation, with one RD child misclassified as ND; $\kappa = .94$). Only 3 variables were used: word reading errors, text reading time, and word spelling errors.

Thus, it seems clear that the unambiguous classifications of the experts are determined on the basis of reading and spelling performance only, taking into account speed as well as accuracy. No other measures seem to influence expert judgment significantly, consistent with the results of the comparison between school and clinical samples (see Table 2).

Note that, because the discriminant analysis was conducted on the same measures used by the experts for their judgment, it is somewhat trivial that high classification rates are obtained. It must be stressed that the point of this analysis was not to examine the classification rates but, rather, to determine which measures influenced expert judgment most, how reliably we can predict expert judgment from a specific set of measures, and, most important, whether the measures apparently taken into account by the experts were the ones that best distinguished the populations (by reference to the analysis of group differences between the school and clinical sample). This is also why we used stepwise analysis, in which a statistical criterion—rather than researchers' judgment—determines which variables enter the discriminant function. The results indicate that—in this case at least—expert judgment for unambiguous cases is quite systematic and reliable with respect to the observed data

patterns. As one might expect, for the cases of initial disagreement, the reliance on the assessment measures appears less systematic, and the corresponding classification performance is not as high (κ range = .70–.77).

Psychometric Adequacy of Computer-Based Assessment

Having established a dichotomous reference criterion based on expert judgment, which is itself valid with respect to the measures found to distinguish the RD-rich sample from the general population, we can calculate the accuracy of matching this criterion with measures from the computer-based assessment. That is, we can examine the validity of computer-based screening in predicting expert judgment regarding the referral of a child for full clinical assessment.

Criterion-Referenced Validity of Screening. Using the 10 variables of the computer-based assessment, with expert judgment serving as the classification standard, stepwise discriminant analysis resulted in a function correctly classifying 97% of the 114 unambiguously classified children (also in cross-validation), with 1 RD group child misclassified as ND and 2 ND group children misclassified as RD (κ = .83). Only 4 variables were used: text reading time, text spelling correction time, pseudoword dictation score, and word–picture matching score.

Further discriminant analyses were conducted with data from all 134 children, using the agreed-upon classification (23 RD total), which resulted in 85% of cases correctly classified (84% cross-validated: 16 RD misclassified as ND, and 5 ND misclassified as RD; κ = .32) on the basis of only two measures: text reading time and text spelling correction errors. Note that this is a very difficult test for computer-based assessment, because the criteria for expert judgment are not perfectly systematic in the case of disagreement between the two experts. Nevertheless, the "performance" of the automated screening procedure was comparable to the interrater reliability. Because expert judgment was the reference criterion, it cannot possibly be expected that a higher rate of correct classification be obtained from the software.

A final analysis was conducted with a manually selected subset of the variables, taking into account (a) the correlations between the measures, (b) their power in discriminating the school sample from the clinical sample, and (c) the time needed for the children to complete each task. Specifically, the two measures from the lengthy spell-checking task were forced out of the analysis, because the much briefer word–picture matching task apparently provided about the same information. Text comprehension score was also kept out of the analysis, along with the three auditory measures, none of which contributed to the discrimination of the samples. The remaining four variables were all entered together into a new discriminant analysis, and the resulting function achieved 97% correct classification (96% cross-validated; κ = .76) of the unambiguous cases. As there was no reduction in classification accuracy over the previous functions, the outcome of this discriminant analysis was retained in further analyses and will henceforth be considered *the* discriminant function for the automatic classification of children as RD or ND in the computer-based assessment. The statistical evaluation for this function indicated an eigenvalue of .48, canonical correlation .57, and Wilks' $\lambda = .675$, $\chi^2(4) = 43.2$, $p < .0005$. Function values at group centroids were –0.20 for ND and 2.35 for RD. Table 6 lists the parameters indicating the relative importance and significance of each variable used in the resulting function (all significant to $p < .0005$).

In conclusion, it is possible to obtain highly valid classifications of the children based on only half of the computer-based assessment, which are comparable in validity (bottom-line classification performance) with the interrater reliability of the experts (see Note 3). The measures that were most useful for the classification were the ones that were earlier found to differ most between the school and the clinical samples. This is an important verification of their significance, because the discriminant analysis included only children from the school sample, referenced to expert judgment on the basis of their traditional assessment.

Test–Retest Reliability of Computer-Based Screening. Once the criterion validity of the computer-based assessment is established in detecting the children most likely to have RD, it is important to also determine the reliability of this detection. The reliability of computer-based screening can be assessed using the standard test–retest

TABLE 6
Statistical Parameters and Relative Importance of Computer-Based Assessment Variables Selected for the Discriminant Function to Classify Children as Having RD or ND

Variable	Wilks' λ	Coefficient	Loading
Text reading time	.829	.565	.716
Pseudoword dictation	.886	.389	.654
Word–picture matching	.802	.568	.516
Letter span	.930	.059	.396

Note. Wilks' lambda, standardized canonical discriminant function coefficients, and discriminant loadings (pooled within-group correlations between discriminant variables and the standardized canonical discriminant function). RD = reading disabilities; ND = no disabilities.

procedure, which evaluates the stability of outcomes for each child. The retest sample (including 48 children who did not contribute data to any other analyses), instead of receiving both types of assessments, completed the computer-based assessment twice.

The correlations between the first and second score obtained on each of the variables for the retest sample are shown in Table 7. Most of the measures appear to be moderately to highly reliable, with spell-checking time being the least reliable and spell-checking errors (number of remaining errors after corrections) being the most reliable. The reliability of these measures is apparently unrelated to their validity for discriminating between the school and clinical sample or for classifying children into RD and ND groups with respect to expert judgment. Auditory measures, in particular, are quite reliable in retest, yet they are useless in the discriminations relevant to RD for this population.

The test–retest reliability of the discriminant function, computed as the Pearson product-moment correlation among the function values derived from the first and second administration of the computer-based assessment, was $r = .57$, $p < .0005$. The reliability of classification itself can be determined by cross-tabulating the RD/ND outcomes from each administration of the computer-based assessment. Table 8 shows that more than half of the children classified as having RD in the first administration were reclassified as having RD in the second administration. The percentage of overall consistent classification in this case is 83% ($\kappa = .49$). This performance is comparable to the interrater reliability, against which the discriminant function was computed, and about as much as one would expect from the criterion validity of the screening as computed on the full sample (including ambiguous cases, which can be expected to occur in the retest sample as well).

However, the important question concerns the children identified as having RD only once: Were they inappropriately missed the other time? Or did they not have RD and only showed up once because of the moderately low specificity of the screening? In other words, if the screening procedure systematically detects unambiguous RD cases with high validity, but also incorrectly detects ND cases less consistently, the actual clinical significance of the screening may be even higher than 83%—that is, *more* reliable than the human experts' judgment based on the measures of the traditional assessment. To examine this question, more detailed validity analyses are needed.

Validity Confirmation of Computer-Based Screening. For the 79 children (37 girls) whose data were not used in the expert classification and the derivation of discriminant functions, it is possible to examine the validity of computer-based screening by comparing the classification derived from the software measures to that derived by traditional assessment. Because a small subset of the traditional assessment measures was found to predict expert judgment almost 100% for unambiguous cases, it was not considered necessary to examine the remaining individual assessment sheets and manually classify the children as having RD or ND. Applying the discriminant function derived from the traditional assessment analysis of only unambiguous cases classified 35 children as RD (11 from the school sample)

TABLE 7
Correlations Between Individual Transformed Measures from First and Second Administration of Computer-Based Assessment in Retest Sample

Measure	*r*	*p*
Text comprehension score	.33	.027
Text reading time	.24	.11
Text spell-checking errors	.80	< .0005
Text spell-checking time	.15	.32
Frequency discrimination	.49	< .0005
Two-tone sequencing	.78	< .0005
Three-tone sequencing	.52	< .0005
Pseudoword dictation score	.65	< .0005
Word-picture matching score	.50	< .0005
Letter span score	.31	.035

Note. Retest sample, $n = 48$.

TABLE 8
Classification of Retest Sample as RD or ND Based on First and Second Administration of the Computer-Based Assessment

	Classification from 2nd assessment	
Classification from 1st assessment	**ND**	**RD**
ND	34	4
RD	4	6

Note. RD = reading disabilities; ND = no disabilities.

and 44 as ND (40 from the school sample). Based on the results of the criterion validity analyses, we expect this classification to agree about 90% with expert judgment (because ambiguous cases are likely to occur in this group of children as well).

Table 9 shows the number of children classified as having RD by the software measures for each category. Results are shown separately for the full sample and the school sample only, because the latter is expected to match, in correct classification probability, the sample on which the discriminant functions were derived. The concordance rates for these classifications are $\kappa = .59$ (all) and $\kappa = .48$ (school only). In both cases, "correct" classification is about 80%, as would be expected from the preceding validity analyses given the imperfect reference criterion.

Any differences arising from adjusting the discrimination threshold will affect the balance between sensitivity (the percentage of RD children correctly classified by the software) and positive predictive value (PPV; the percentage of children classified as RD by the software that are correctly classified) of the computer-based screening. By altering the critical cutoff value of the discriminant function, it is thus possible to bias the selection toward maximal sensitivity (i.e., to detect as many RD as possible), in which case PPV may drop to 50% or lower.

If the computer-based screening is successful in detecting the children who are most likely to be diagnosed with RD, then the discrimination criterion (cutoff threshold on the discriminant function) should allow adjustment of certainty (with which each child is referred) against coverage (successful detection of most RD children). A useful discriminant function would identify the children with the most severe RD (and only children with RD) at the most stringent criterion. As the criterion is set to lower threshold values, the detection of a higher proportion of children with RD will be necessarily accompanied by an increasing number of ND children misidentified as having RD. Therefore, the derived discriminant function will be clinically useful to the extent that its positive predictive value increases as its sensitivity is decreased.

Figure 1 shows the sensitivity and positive predictive value of computer-based screening as the number of detected children increases (expressed as a ratio over the number of target children with RD, to allow comparisons across different groups of children, some including the clinical sample). It is evident that the relation between sensitivity and PPV is strongly inverse, as expected. Moreover, the balance between sensitivity and PPV is not affected very much by changes in the proportion of children with RD in the population or in the strictness of RD classification criteria. Consistent with the findings from the discriminant analysis sample, the classification as (possibly) having RD of twice the number of children actually expected to have RD results in the detection of about 90% of the children with RD in the population.

Therefore, the bottom-line total correct classification by the software is comparable to the human interrater reliability in the traditional assessment and remains in agreement with the corresponding result of the initial discriminant analysis when considering all children, including those whose classification was initially in disagreement by the experts. The uncertainty in the classification of the 20 children in the initially ambiguous group is seen clearly in the failure of the discriminant function based on the traditional assessment measures to match the experts' agreement for 12 of them, classifying as having RD only 6 of the 14 children indicated by the two experts after discussion. Considering that the same function accounted for 100% of the unambiguous cases, and that interrater reliability was 85%, it seems more likely that the difficulty lies with the profiles of the children and with the lack of common standards and definitions in the Greek system than with any inadequacy of the statistical analyses.

Computer-Based Detection of Slow and Inaccurate Readers. Because dyslexia is typically understood as a deficit that primarily affects word-level reading accuracy and fluency, it is of interest to examine the ability of computer-based measures to separate groups formed on the basis of single-

TABLE 9
Classification of Children from the School and Clinical Samples as RD or ND Based on the Traditional Versus the Computer-Based Assessment

	Classification from computer-based assessment	
Classification from traditional assessment	**ND**	**RD**
Combined group[a]		
ND	34	10
RD	6	29
School sample only[b]		
ND	31	9
RD	2	9

Note. RD = reading disabilities; ND = no disabilities.
[a]Includes all 79 children from the school and clinical samples not included in the calculation of the discriminant functions. [b]Includes only the 51 children from the school sample not included in the calculation of the discriminant functions.

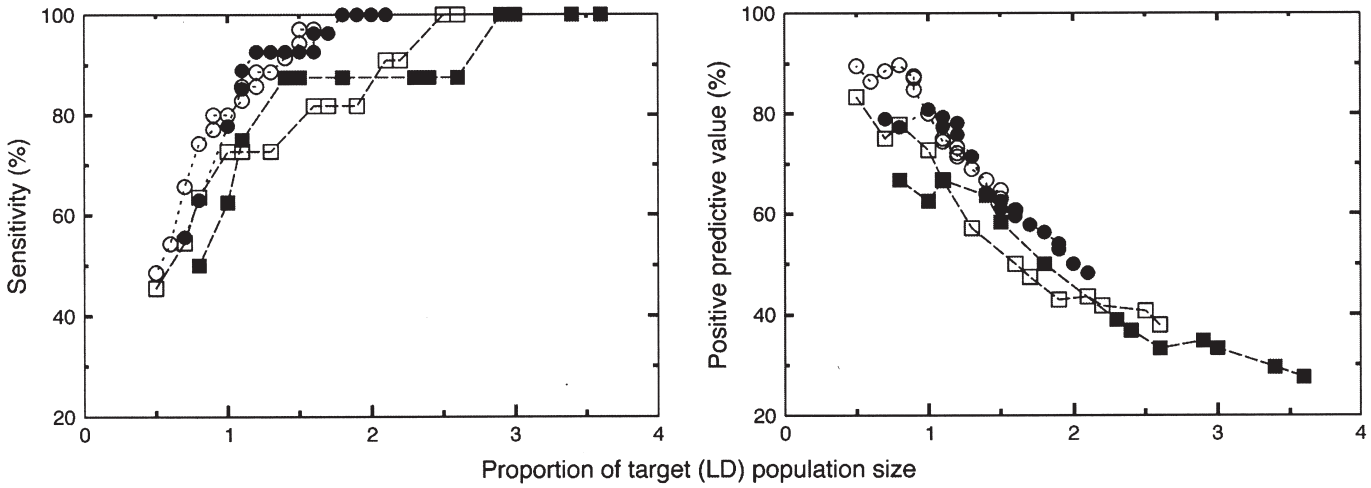

FIGURE 1. Computer-based detection sensitivity and positive predictive value as a function of the proportion of children identified as having RD (that is, the ratio of children falling in the "RD" range of the discriminant function value over the number of children assumed to have RD by the reference criterion). For example, if 10% of the sample is "known" to have RD, and 20% of the children are classified as having RD by the software, this would correspond to a value of 2.0 on the horizontal axis. The proportion of the "true" 10% that are classified as having RD is shown in the left panel and the proportion of the classified 20% that are "true" RD is shown on the right panel. (The calculation was meant to allow comparisons across populations with different proportions of RD children). The proportion of children identified as having RD is adjusted by altering the critical cutoff value of the discriminant function that distinguishes RD from ND classification. For these graphs, cutoff values range between 0.0 and 2.0. Circles = all children; Squares = school sample only; Outline markers = liberal classification from traditional assessment (cutoff = 2.0); Filled markers = Conservative classification from traditional assessment (cutoff = 2.5).

word reading speed and accuracy. High- and low-performing groups on these two measures from the traditional assessment were formed as follows: For word reading time, scores up to the 20th percentile (44 children) defined one group, and scores from the 39th percentile up (130 children) defined the second group (henceforth *speed grouping*). For word reading errors, scores up to the 21st percentile (46 children) defined one group, and scores from the 39th percentile up (131 children) defined the second group (henceforth *accuracy grouping*).

Table 10 shows that the computer-based measures, with the exception of text comprehension and the auditory tasks, differed significantly between the high- and low-performing groups for both the speed and the accuracy grouping. After excluding the variables not differing significantly, linear discriminant function analyses were conducted separately for each computer-based measure alone and for all six measures together (see Table 11). Text reading speed was the best predictor of the speed grouping, followed by spell-checking time, whereas spell-checking errors was the best predictor of the accuracy grouping, followed by word–picture matching. Thus, consistent with the dimensions found in the factor analyses, two speed measures best predicted word reading speed, and two spelling measures best predicted reading accuracy. Naturally, the inclusion of all computer-based variables in the discriminant function increases the detection performance, because of the intercorrelations and of the imperfect reliability of the measures.

This analysis should not be taken to imply that a classification of RD can be made on the basis of a single measure. A single measure cannot capture all the relevant skill variance associated with a target dimension (e.g., accuracy). Moreover, imperfect reliability and the lack of educational and clinical context dictate that such groupings can only be indicative. Still, it is of interest that these theoretically fundamental single-variable groupings can be significantly approximated with the automated computer-based measures, at a level of concordance comparable to the validity of screening.

Discussion

The purpose of this study was to examine the characteristics of the RD population in Greece and draw conclusions for further study related to RD assessment practice, both traditional and computer based. Age was held constant in our sample to ensure the reliability of comparisons among measures and to avoid complexities related to development or amount of instruction. However, the lack of age variability also means that any conclusions drawn from this sample will be of questionable generalizability to other

TABLE 10
Analysis of Variance of Computer-Based Assessment Measures for Speed and Accuracy Groupings of Children Based on Traditional Assessment

	Accuracy grouping[a]				Speed grouping[b]			
Computer-based measure	***F*(1, 175)**	***p***	η^2	***d***	***F*(1, 175)**	***p***	η^2	***d***
Text comprehension score	1.625	.204	.009	.219	3.178	.076	.018	.311
Text reading time	36.505	< .0005	.173	1.034	80.825	< .0005	.320	1.567
Text spell-checking errors	61.811	< .0005	.261	1.348	68.666	< .0005	.285	1.446
Text spell-checking time	13.216	< .0005	.070	.623	44.822	< .0005	.207	1.167
Frequency discrimination	0.804	.371	.005	–.154	2.429	.121	.014	–.272
Two-tone sequencing	0.924	.338	.005	–.165	0.150	.699	.001	.067
Three-tone sequencing	5.576	.019	.031	–.405	1.254	.264	.007	–.195
Pseudoword dictation score	31.853	< .0005	.154	.967	14.959	< .0005	.080	.674
Word–picture matching score	62.534	< .0005	.263	1.356	64.062	< .0005	.271	1.396
Letter span score	20.475	< .0005	.105	.776	33.971	< .0005	.165	1.017

Note. *d* = Cohen's measure of normalized distance among group means.
[a]Scores up to the 20th percentile versus 39th percentile and higher on word reading errors. [b]Scores up to the 21st percentile versus 39th percentile and higher on word reading time.

TABLE 11
Linear Discriminant Function Analysis Using Individual Computer-Based Assessment Measures as Predictors of Speed and Accuracy Groupings of Children Based on Traditional Assessment

	Accuracy grouping[a]				Speed grouping[b]			
Computer-based measure	**% Corr**	**Sens**	**PPV**	κ	**% Corr**	**Sens**	**PPV**	κ
Text reading time	70.1	.717	.452	.346	78.2	.750	.550	.484
Text spell-checking errors	78.5	.783	.563	.505	73.6	.773	.486	.415
Text spell-checking time	63.8	.609	.378	.215	75.9	.773	.515	.452
Pseudoword dictation score	68.9	.739	.442	.337	66.1	.659	.397	.263
Word–picture matching score	71.8	.783	.474	.394	72.4	.773	.472	.397
Letter span score	63.3	.739	.391	.260	70.7	.818	.456	.386
All six entered together	79.1	.739	.576	.502	83.9	.773	.654	.598

Note. % Corr = percentage correctly classified; Sens = sensitivity; PPV = positive predictive value; κ = Cohen's measure of concordance among groupings.
[a]Scores up to the 20th percentile versus 39th percentile and higher on word reading errors. [b]Scores up to the 21st percentile versus 39th percentile and higher on word reading time.

ages. Therefore, the following discussion applies only to seventh-grade children.

Assessment Structure and the Characteristics of RD in Greek

The analyses showed that the clinical sample—a population with a very large proportion of children with RD—is distinguished from the general population primarily on the basis of slow reading, consistent with reports from other languages with regular orthography (reviewed in the introduction) and also from much younger children learning to read Greek (Porpodas, 1999). Reading and spelling accuracy problems are also of substantial importance. These findings are important and reliable because they emerged independently in the self-selected clinical sample, in the experts' judgment of the school sample only, and in the factor structure of the assessment battery.

The traditional assessment seemed to address mainly three dimensions of relevant skills: accuracy, fluency, and intelligence. None of these factors sep-

arates reading from spelling ability, indicating that orthographic precision and efficiency affect the expressive and receptive aspects of written communication equally. In contrast, the data suggest a distinction between accuracy and fluency, even though several measures contribute to both dimensions. Critically, timing measures (including color naming) contributed only to the fluency dimension, whereas phoneme deletion—our only phonological awareness measure—contributed only to the accuracy dimension, and not to fluency. These results are consistent with a two-factor model for reading and spelling skills, in agreement with the double-deficit hypothesis (Wolf & Bowers, 1999; Wolf et al., 2002) and with corresponding data from German (Wimmer, Mayringer, & Landerl, 2000). The relatively small effect size for phoneme deletion indicates that by this age, core phonological difficulties have to a large extent been resolved and are thus not strong indicators of RD.

Our findings may be taken as dissimilar to the German findings of Wimmer and Mayringer (2002), who found dissociations between reading and spelling difficulties, possibly reflecting a sample difference, because the children in that German study were much younger (Grades 3 and 4). However, there may actually be no discrepancy, as the dissociation in German between reading fluency and spelling parallels the dissociation in Greek between accuracy and fluency. Wimmer and Mayringer did not report reading error analyses because of performance "close to ceiling" (p. 273); they found a dissociation between spelling accuracy, related to phonological awareness, and reading fluency, related to naming speed. Word reading accuracy was also very high in our sample: The median for the school sample was one error, and the median for the clinical sample was 10% errors, equal to the 90th percentile of the school sample. Moreover, our phonological awareness measure loaded on the accuracy factor, along with word reading and spelling accuracy, whereas speed measures loaded on the fluency factor. Thus, if the few errors made by the German children were analyzed and found to be strongly related to spelling accuracy, then our findings would replicate those of Wimmer and Mayringer.

The emergence of three factors in the analysis of the assessment battery does not necessarily mean that these factors are the only relevant dimensions or even that these dimensions are all related to RD. Because the sample is composed primarily of children with no RD, the observed factor structure may reflect the "typical" state of affairs as far as skill domains are concerned. If we were to examine only children with RD, it is conceivable that due to the RD, the intercorrelations among the measures might be lower (or differently patterned) and, therefore, that a different structure might emerge. Unfortunately, the available sample of children with RD was not large enough to permit independent analyses.

The relevance of the three factors for RD assessment can be examined by ANOVA comparing the school and the clinical samples on regressed factor scores. Significant differences were found for Factor 1 (accuracy), $F(1, 211) = 15.75$, $p < .0005$, $\eta^2 = .07$; and Factor 2 (fluency), $F(1, 211) = 53.69$, $p < .0005$, $\eta^2 = .20$; but not for Factor 3 (intelligence), $F(1, 211) < 1$. Therefore, even though some of the measures loading on Factor 3 did differ significantly between the two samples, it appears that their common variance expressed in the intelligence dimension is not relevant for the distinction between children with and without RD. This is consistent with the current trend away from intelligence measures in RD assessment. The larger effect size for Factor 2 is also consistent with the greater importance placed on fluency than on accuracy, especially for orthographically regular languages.

Because word-level difficulties, and not text comprehension, are considered to be the defining feature of dyslexia, the finding that the RD-rich clinical sample does not differ very much from the general population on the reading comprehension measures is in agreement with current understanding of RD, in that reading comprehension is not in itself a typical problem for Greek children with RD. In saying that, we must not draw the conclusion that difficulty with comprehension is not frequently a sign of LD. In our study, the reading comprehension subtest consisted of three passages, the two most demanding of which were read silently for 1 minute after the initial oral reading. This may have diminished differences in comprehension, insofar as poor readers had sufficient time to compensate for any working memory or text processing deficits, by expressing a tradeoff between the time taken to read (measured as fluency) and text processing time (primarily affecting comprehension; see Weaver, 2000). It remains possible that a subgroup of children with LD present with primary difficulties in text comprehension. Such children should not be diagnosed with dyslexia, and, on the basis of the aforementioned findings, it might be expected that their performance on the intelligence measures would be on the low side of average (Oakhill & Garnham, 1988).

The Case for Computer-Based Screening

Because of the lack of a school-based service system for assessing and helping children with LD in Greece, children with learning problems manage to reach secondary school unable to deal with the demands of academic work. Few professionals are adequately trained to deal with assessment and intervention in an educationally constructive manner, and so far no standardized screening battery is available in the Greek language for the age group studied here. In this context, computer-based assessment may prove instrumental in the quest for well-researched psychometric tools. Current diagnostic practice is based primarily on personal experience, rely-

ing more on overall impression than on specific measurement, and thus less likely to identify the pattern of difficulties that each student faces and less likely to devise the proper intervention addressing the domains of greatest need. Clinical instinct may be one important factor, but it must be complemented with psychometrically validated tools. After all, clinical judgments are known to be inferior to actuarial classifications, especially when based on unstructured observations and not on standardized measures (Grove, Zald, Lebow, Snitz, & Nelson, 2000).

If computer-based assessment is to be employed as a screening tool, it must be held to the usual psychometric standards of traditional measures (cf. Naglieri, 2004, for Internet-based testing). As other recent attempts have indicated, this is a feasible aim (e.g., Lancaster & Mellard, 2005). Normative data and high validity and reliability are necessary, as well as clear implications for application in educational practice. The battery presented here has only passed the first stage of psychometric adequacy, in a relatively small sample of the general school population. Large-scale application and normative data are needed before its widespread adoption.

Computer-based assessment can never supplant evaluation by an expert professional, primarily because it cannot provide the comprehensive evaluation needed to secure proper diagnosis. Current recommendations for assessments dictate "a hierarchical approach . . . in which the relationship of academic deficits, cognitive skills, and psychosocial factors is carefully evaluated for each child [to allow] an interdisciplinary team to develop an intervention plan" (Fletcher et al., 2002, p. 57). This feat is—and will likely remain—outside the reach of computer-based assessment. However, if the specific measures generated by computer-based assessment are reliable, valid, and useful, why not import the technology into standard practice as a *supporting* tool?

Financial and social considerations are also relevant: If there are insufficient personnel to address the needs of an educational system, this means that many children who need extra attention will simply not receive it. Upon this null baseline, *any* method that can offer reliable (even if imperfect) services to these children at an acceptable cost should be taken seriously. As all schools now have some computer equipment, a computer-based screening test comes at negligible cost (and with none of the stigma associated with visits to special services). Insofar as this test can identify a subgroup of children who are most likely to need extra attention and refer them for evaluation, the cost–benefit advantages of this approach are obvious.

Conclusion

Turning to the research questions stated in the introduction, the following tentative answers may be provided on the basis of our findings:

1. The population of schoolchildren in Greece who seek assessment for LD and are likely to receive a diagnosis of dyslexia according to current practice differ from the general school population primarily in reading speed, and also in reading and spelling accuracy.
2. The factor structure of the assessment battery includes three main dimensions, related to accuracy, fluency, and intelligence. Of these, only the first two are relevant for the RD diagnosis.
3. Expert judgment for RD diagnosis, at least for the two professionals who contributed to this study, seems to be based primarily on reading and spelling accuracy and on reading speed (i.e., on the same measures that distinguish the relevant population). On the other hand, discrepant judgments were also noted, which are not accounted for statistically in a consistent manner and thus remain in need of further investigation.
4. Computer-based screening is psychometrically adequate in detecting children who are likely to be diagnosed with RD if evaluated by competent professionals. A valid and reliable discriminant function can be derived from measures taken without supervision in a 30-min interaction in the form of "computer games" in the school computer laboratory. The test–retest reliability and the criterion-referenced validity of the detection are comparable to the interrater reliability among independent experts. The sensitivity and predictive value of detection can be balanced against each other to achieve optimal cost–benefit results for a specified proportion of children with RD in the general population.

Future research should address in detail the characteristics of children with RD using a more comprehensive battery and a larger RD population, to determine whether the subtypes reported for English are also relevant for Greek, with the ultimate objective of prescribing intervention practices tailored to the individual needs of children with RD according to their performance patterns.

ABOUT THE AUTHORS

Athanassios Protopapas, *PhD, is a principal researcher at the Institute for Language and Speech Processing in Athens, Greece. His research interests include speech perception and phonetic training, reading and educational assessment, and brain function and connectionist modeling of oral and written language.* ***Christos Skaloumbakas,*** *MEd, is a learning disabilities specialist at the Athens Medical Pedagogical Center and a doctoral candidate in special education at the University of Birmingham in England. He works on assessment and remediation of LD in elementary and secondary education, with particular interest in the development of written expression and spelling. Address: Athanassios Protopapas, ILSP, Artemidos 6 & Epidavrou, GR-15125 Maroussi, Greece; e-mail: protopap@ilsp.gr*

AUTHORS' NOTES

1. *This research was originally made possible in part by a contract from the Hellenic Pedagogical Institute (PI) to the Institute for Language and Speech Processing (ILSP) to develop screening software for seventh-grade students with LD, in the context of Ministry of Education program "EPEAEK" actions 1.1.β and 1.4.γ ("Greek school network"). The software was developed by ILSP programmers D. Routsis and G. Koulafetis, with graphic design by G. Magakis and A. Glarou.*
2. *The authors are indebted to PI members G. Papadopoulos, M. Karamanis, A. Kriba, and V. Ioannou for their help with the software trials; to P. Vrondos and M. Iliopoulou for making the test–retest trial posssible; to the principals and teachers of the participating schools for accepting and facilitating the testing; to the parents and children for their participation; and to E. Stamou, A. Archonti, T. Triandafyllakos, C. Tsagaraki, A. Grigoriadou, S. Gerakaki, and S. Alexandri for administering the assessment tests. Many thanks are due to D. Nikolopoulos for significant help in the early stages of the design and application of the testing batteries and to S. Georgakakou and Y. Vogindroukas for helpful discussions. We are grateful to K. Mylonas for help with the statistical analyses and to G. Sideridis, I. Dimakos, V. Kourbetis, and A. Mouzaki for comments on the manuscript.*
3. *Partial preliminary results of the studies reported here have been reported in Greek conferences between 2001 and 2004.*

NOTES

1. *There was also a gender difference in frequency discrimination, $F(1, 208) = 10.03$, $p = .002$, interacting with sample, $F(1, 208) = 5.22$, $p = .023$, so that it was significant only in the clinical sample, with boys performing better than girls, $F(1, 25) = 23.99$, $p < .0005$. Because this measure was not found to be relevant to the RD classification, the difference was not pursued further.*
2. *This second factor is also the one on which a naming measure would load most. A rapid color naming measure was included in the traditional assessment battery, but it was removed from the analysis because it was not part of the original protocol and data for only 162 children of the 213 were available. For these 162 children, Pearson's correlation coefficients (r) between time to name 60 colored "XXXXX" items and the three factors were −.12, .49, and −.26. Of these, only the correlation with Factor 2 (fluency) was significant to $p < .0005$.*
3. *To verify that the small size of the RD group was not the cause of the highly successful discrimination, additional analyses were conducted including the clinical sample. The new ND group thus formed included the 105 children from the school sample plus 2 children from the clinical sample who were not diagnosed with RD, and the new RD group included the 9 children from the school sample who were unambiguously judged to have RD by both experts plus 26 children from the clinical sample who were diagnosed with RD after comprehensive clinical evaluation. The resulting discriminant functions were practically identical with those derived using only the 9 RD children from the school sample: Correlation coefficients among discriminant scores from the two analyses were $r = .99$ for the traditional assessment and $r = .98$ for the computer-based assessment.*

REFERENCES

Adlard, A., & Hazan, V. (1998). Speech perception in children with specific reading difficulties. *The Quarterly Journal of Experimental Psychology, 51A,* 153–177.

Ahissar, M., Protopapas, A., Reid, M., & Merzenich, M. M. (2000). Auditory processing parallels reading abilities in adults. *Proceedings of the National Academy of Sciences of the United States of America, 97,* 13907–13912.

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.

Beitchman, J. H., & Young, A. (1997). Learning disorders with a special emphasis on reading disorders: A review of the past 10 years. *Journal of the American Academy of Child and Adolescent Psychiatry, 36,* 1020–1032.

Bishop, D. V. M., North, T., & Donlan, C. (1996). Nonword repetition as a behavioural marker for inherited language impairment: Evidence from a twin study. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 37,* 391–403.

Bruck, M. (1992). Persistence of dyslexics' phonological awareness deficits. *Developmental Psychology, 28,* 874–886.

Caravolas, M., & Volín, J. (2001). Phonological spelling errors among dyslexic children learning a transparent orthography: The case of Czech. *Dyslexia, 7,* 229–245.

Chitiri, H., & Willows, D. (1994). Word recognition in two languages and orthographies: English and Greek. *Memory & Cognition, 22,* 313–325.

Chliounaki, K., & Bryant, P. (2002). Construction and learning to spell. *Cognitive Development, 17,* 1489–1499.

Cossu, G., Shankweiler, D., Liberman, I. Y., Katz, L., & Tola, G. (1988). Awareness of phonological segments and reading ability in Italian children. *Applied Psycholinguistics, 9,* 1–16.

De Weirdt, W. (1988). Speech perception and frequency discrimination in good and poor readers. *Applied Psycholinguistics, 9,* 163–183.

Defior, S., Martos, F., & Cary, L. (2002). Differences in reading acquisition development in two shallow orthographies: Portuguese and Spanish. *Applied Psycholinguistics, 23,* 135–148.

Ellis, N. C., Natsume, M., Stavropoulou, K., Hoxhallari, L., van Daal, V., Polyzoe, N., et al. (2004). The effects of orthographic depth on learning to read alphabetic, syllabic, and logographic scripts. *Reading Research Quarterly, 4,* 438–468.

Fletcher, J. M., Foorman, B. R., Boudousquie, A., Barnes, M. A., Schatschneider, C., & Francis, D. J. (2002). Assessment of reading and learning disabilities: A research-based intervention-oriented approach. *Journal of School Psychology, 40,* 27–63.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5,* 239–256.

Gathercole, S. E., Willis, C., Baddeley, A. D., & Emslie, H. (1994). The children's test of nonword repetition: A test of phonological working memory. *Memory, 2,* 103–127.

Georgas, D. D., Paraskevopoulos, I. N., Bezevegis, I. G., & Giannitsas, N. D. (1997). *Ελληνικό WISC-III*: *Wechsler κλίμακες νοημοσύνης για παιδιά* [Greek WISC-III: Wechsler intelligence scale for children]. Athens: Ellinika Grammata.

Gottardo, A., Siegel, L. S., & Stanovich, K. E. (1997). The assessment of adults with reading disabilities: What can we learn from experimental tasks? *Journal of Research in Reading, 20,* 42–54.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-

analysis. *Psychological Assessment, 12,* 19–30.

Hannon, B., & Daneman, M. (2001). A new tool for measuring and understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology, 93,* 103–128.

Hatzigeorgiu, N., Gavrilidou, M., Piperidis, S., Carayannis, G., Papakostopoulou, A., Spiliotopoulou, A., et al. (2000, May 31–June 2). Design and implementation of the online ILSP corpus. In *Proceedings of the second international conference of language resources and evaluation* (Vol. 3, pp. 1737–1740). Athens, Greece.

Holopainen, L., Ahonen, T., & Lyytinen, H. (2001). Predicting delay in reading achievement in a highly transparent language. *Journal of Learning Disabilities, 34*, 401–413.

Ives, B. (2003). Effect size use in studies of learning disabilities. *Journal of Learning Disabilities, 36,* 490–504.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Accuracy and fluency in list and context reading of skilled and RD groups: Absolute and relative performance levels. *Learning Disabilities Research & Practice, 18*, 237–245.

Jiménez González, J. E., & Hernández Valle, I. (2000). Word identification and reading disorders in the Spanish language. *Journal of Learning Disabilities, 33*, 44–60.

Lancaster, S., & Mellard, D. (2005). Adult learning disabilities screening with an Internet-administered instrument. *Learning Disabilities: A Contemporary Journal, 3, 62*–73.

Landerl, K. (2001). Word recognition deficits in German: More evidence from a representative sample. *Dyslexia, 7,* 183–196.

Landerl, K., & Wimmer, H. (2000). Deficits in phoneme segmentation are not the core problem of dyslexia: Evidence from German and English children. *Applied Psycholinguistics, 21,* 243–262.

Landerl, K., Wimmer, H., & Frith, U. (1997). The impact of orthographic consistency on dyslexia: A German–English comparison. *Cognition, 63,* 315–334.

Lehtola, R., & Lehto, J. E. (2000). Assessing dyslexia in Finnish high school students: A pilot study. *European Journal of Special Needs Education, 15,* 255–263.

Lyon, G. R., Fletcher, J. M., & Barnes, M. C. (2002). Learning disabilities. In E. J. Mash & R. Barkley (Eds.), *Child psychopathology* (2nd ed., pp. 520–586). New York: Guilford Press.

Maridaki-Kassotaki, A. (1998). Ικανότητα βραχύχρονης συγκράτησης φωνολογικών πληροφοριών και επίδοση στην ανάγνωση: Μια προσπάθεια διερεύνησης της μεταξύ τους σχέσης [Evaluation of the relationship between phonological working memory and reading ability in Greek-speaking children]. *Psychologia, 5,* 44–52.

Mavrommati, T. D., & Miles, T. R. (2002). A pictographic method for teaching spelling to Greek dyslexic children. *Dyslexia, 8,* 86–101.

Mayringer, J., & Wimmer, H. (2000). Pseudoname learning by German-speaking children with dyslexia: Evidence for a phonological learning deficit. *Journal of Experimental Child Psychology, 75,* 116–133.

McBride-Chang, C. (1995). What is phonological awareness? *Journal of Educational Psychology, 87,* 179–192.

Muller, K., & Brady, S. (2001). Correlates of early reading performance in a transparent orthography. *Reading and Writing: An Interdisciplinary Journal, 14,* 757–799.

Naglieri, J. A. (2004). Psychological testing on the Internet: New problems, old issues. *The American Psychologist, 59,* 150–162.

Oakhill, J., & Garnham, A. (1988). *Becoming a skilled reader*. Oxford, UK: Blackwell.

Onwuegbuzie, A. J., Levin, J. R., & Leech, N. L.(2003). Do effect-size measures measure up?: A brief assessment. *Learning Disabilities: A Contemporary Journal, 1,* 37–40.

Paraskevopoulos, I. N., Kalantzi-Azizi, A., & Giannitsas, N. D. (1999). *ΑθηνάΤεστ διάγνωσης δυσκολιών μάθησης* [AthenaTest for the diagnosis of learning difficulties]. Athens: Ellinika Grammata.

Porpodas, C. D. (1999). Patterns of phonological and memory processing in beginning readers and spellers of Greek. *Journal of Learning Disabilities, 32,* 406–416.

Rack, J., Snowling, M., & Olson, R. (1992). The nonword reading deficit in developmental dyslexia: A review. *Reading Research Quarterly, 27,* 29–53.

Ramus, F. (2001). Outstanding questions about phonological processing in dyslexia. *Dyslexia, 7,* 197–216.

Ramus, F., Rosen, S., Dakin, S. C., Day, B. L., Castellote, J. M., White, S., et al. (2003). Theories of developmental dyslexia: Insights from a multiple case study of dyslexic adults. *Brain, 126,* 1–25.

Raven, J. (1976). *Standard progressive matrices.* New York: Psychological Corp.

Reed, M. A. (1980). Speech perception and the discrimination of brief auditory cues in reading disabled children. *Journal of Experimental Child Psychology, 48,* 270–292.

Sabatini, J. P. (2002). Efficiency in word reading of adults: Ability group comparisons. *Scientific Studies of Reading, 6*, 267–298.

Schatschneider, C., Carlson, C. D., Francis, D. J., Foorman, B. R., & Fletcher, J. M. (2002). Relationship of rapid automatized naming and phonological awareness in early reading development: Implications for the double-deficit hypothesis. *Journal of Learning Disabilities, 35,* 245–256.

Schulte-Körne, G., Deimel, W., Bartling, J., & Remschmidt, H. (1999). The role of phonological awareness, speech perception, and auditory processing for dyslexia. *European Child & Adolescent Psychiatry, 8,* 260–267.

Serniclaes, W., Sprenger-Charolles, L., Carre, L., & Demonet, J. F. (2001). Perceptual discrimination of speech sounds in developmental dyslexia. *Journal of Speech, Language, and Hearing Research, 44,* 384–399.

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

Shaywitz, S. E. (2003). *Overcoming dyslexia: A new and complete science-based program for reading problems at any level.* New York: Knopf.

Shaywitz, S. E., & Shaywitz, B. A. (2003). Dyslexia (specific reading disability). *Pediatrics in Review, 24,* 147–153.

Sideridis, G. D., Morgan, P. L., Botsas, G., Padeliadu, S., & Fuchs, D. (2006). Predicting LD on the basis of motivation, metacognition, and psychopathology: An ROC analysis. *Journal of Learning Disabilities, 39, 215*–229.

Siegel, L. S. (1989). IQ is irrelevant to the definition of learning disabilities. *Journal of Learning Disabilities, 22,* 469–478.

Siegel, L. S. (2003). IQ-discrepancy definitions and the diagnosis of LD: Introduction to the special issue. *Journal of Learning Disabilities, 36,* 2–3.

Stanovich, K. E. (1988). Explaining the difference between the dyslexic and the garden-variety poor reader: The phonological-core variable-difference model. *Journal of Learning Disabilities, 21,* 590–604.

Stein, N. L. (1984). Η ανάπτυξη της ικανότητας των παιδιών να λένε ιστορίες [The development of children's storytelling ability]. In S. Vosniadou (Ed.), *Keimena ekseliktikis psychologias, tomos a: Glossa*. Athens: Gutenberg.

Tallal, P. (1976). Rapid auditory processing in normal and disordered language development. *Journal of Speech and Hearing Research, 19,* 561–571.

Tallal, P. (1980). Auditory temporal perception, phonics, and reading disabilities in children. *Brain and Language, 9,* 182–198.

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K. S., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities, 34,* 33–58.

Torgesen, J. K., & Houck, D. (1980). Processing deficiencies of learning-disabled children who perform poorly on the digit span test. *Journal of Educational Psychology, 72,* 141–160.

Tressoldi, P. E., Stella, G., & Faggella, M. (2001). The development of reading speed in Italians with dyslexia: A longitudinal study. *Journal of Learning Disabilities, 34,* 414–417.

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research, 35,* 2503–2522.

van der Leij, A., & van Daal, V. H. P. (1999). Automatization aspects of dyslexia: Speed limitations in word identification, sensitivity to increasing task demands, and orthographic compensation. *Journal of Learning Disabilities, 32,* 417–428.

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). The development of reading-related phonological processing abilities: New evidence of bi-directional 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.

Warnke, A. (1999). Reading and spelling disorders: Clinical features and causes. *European Child & Adolescent Psychiatry, 8,* 1–12.

Weaver, S. M.(2000). The efficacy of extended time on tests for postsecondary students with learning disabilities. *Learning Disabilities: A Multidisciplinary Journal, 10,* 47–56.

Wimmer, H., & Mayringer, H. (2002). Dysfluent reading in the absence of spelling difficulties: A specific disability in regular orthographies. *Journal of Educational Psychology, 94,* 272–277.

Wimmer, H., Mayringer, H., & Landerl, K. (2000). The double-deficit hypothesis and difficulties in learning to read a regular orthography. *Journal of Educational Psychology, 92,* 668–680.

Wolf, M., & Bowers, P. G. (1999). The double-deficit hypothesis for the developmental dyslexias. *Journal of Educational Psychology, 91,* 415–438.

Wolf, M., Bowers, P. G., & Biddle, K. (2000). Naming speed processes, timing and reading: A conceptual review. *Journal of Learning Disabilities, 33,* 387–407.

Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading, 5,* 211–239.

Wolf, M., O'Rourke, A. G., Gidney, C., Lovett, M., Cirino, P., & Morris, R. (2002). The second deficit: An investigation of the independence of phonological and naming-speed deficits in developmental dyslexia. *Reading and Writing: An Interdisciplinary Journal, 15,* 43–72.

Zahos, G. I., & Zahos, D. I. (1998). *Δυσλεξία. Αντιμετώπιση–Αποκατάσταση. Οδηγίες εφαρμογής προγράμματος* [Dyslexia: Guide to the application of an intervention–remediation program]. Athens: Author.

Ziegler, J. C., Perry, C., Ma-Wyatt, A., Ladner, D., & Schulte-Körne, G. (2003). Developmental dyslexia in different languages: Language-specific or universal? *Journal of Experimental Child Psychology, 86,* 169–193.

Zoccolotti, P., de Luca, M., Judica, A., Orlandi, M., & Spinelli, D. (1999). Markers of developmental surface dyslexia in a language (Italian) with high grapheme–phoneme correspondence. *Applied Psycholinguistics, 20,* 191–216.

(see Appendix on next page)

APPENDIX

Traditional Assessment Intercorrelations

TABLE A1

Correlations Among Measures from the Traditional Assessment

Measure	PRDT	PREP	WRDE	WRDT	TRDE	TRDT	TCOMP	TSPE	WSPE	PHDL	SPDIS	RAV	DIGSP	ARITH
PRDE	**–.41**	.12	**.64**	**–.46**	**.55**	**–.47**	–.22	**.58**	**.53**	**.50**	**.34**	**.24**	**–.31**	**–.37**
PRDT		.02	**–.48**	**.78**	**–.48**	**.72**	.10	**–.52**	**–.38**	**–.26**	–.07	.04	**.25**	.13
PREP			.12	–.12	.12	–.15	**–.25**	.14	.15	.10	**.37**	.18	–.22	**–.25**
WRDE				**–.60**	**.64**	**–.65**	**–.31**	**.66**	**.66**	**.47**	**.31**	.22	**–.34**	**–.31**
WRDT					**–.60**	**.83**	**.27**	**–.64**	**–.56**	**–.39**	–.22	–.07	**.38**	**.27**
TRDE						**–.57**	–.22	**.57**	**.58**	**.47**	**.25**	.20	**–.31**	**–.31**
TRDT							**.37**	**–.71**	**–.62**	**–.40**	**–.26**	–.09	**.43**	**.32**
TCOMP								**–.33**	**–.34**	**–.31**	**–.37**	**–.35**	**.30**	**.38**
TSPE									**.78**	**.47**	**.34**	**.27**	**–.34**	**–.39**
WSPE										**.49**	**.38**	**.30**	**–.32**	**–.41**
PHDL											**.39**	**.36**	**–.42**	**–.39**
SPDIS												**.35**	**–.34**	**–.36**
RAV													**–.25**	**–.41**
DIGSP														**.45**

Note. Pearson product–moment correlation coefficients (*r*) of transformed measures from the traditional assessment, for the school and clinical samples combined ($N = 213$). Correlations significant to $p < .0005$ are shown in bold. PRDE = pseudoword reading errors; PRDT = pseudoword reading time; PREP = pseudoword repetition errors; WRDE = word reading errors; WRDT = word reading time; TRDE = text reading errors; TRDT = text reading time; TCOMP = text comprehension score; TSPE = text spelling errors; WSPE = word spelling errors; PHDL = phoneme deletion errors; SPDIS = speech discrimination errors; RAV = *Raven's Standard Progressive Matrices* (Raven, 1976) raw score; DIGSP = *Wechsler Intelligence Scale for Children,* 3rd ed., Greek version (WISC-III; Georgas et al., 1997) Digit Span raw score; ARITH = WISC-III Arithmetic raw score.

TABLE A2

Correlations Between Measures from the Traditional Assessment (Rows) and Computer-Based Assessment (Columns)

Traditional measure	TCS	TRT	TSE	TST	FD	2TS	3TS	PD	WPM	LS
Pseudoword reading errors	.09	**.27**	**.52**	.17	–.09	–.17	–.21	**.26**	**.48**	**.26**
Pseudoword reading time	–.13	**–.48**	**–.40**	**–.33**	.02	–.05	.02	–.13	**–.42**	**–.25**
Pseudoword repetition errors	.11	.08	.16	.14	–.22	–.21	–.19	.19	.15	.22
Word reading errors	.16	**.44**	**.62**	**.28**	–.07	–.12	–.15	**.35**	**.58**	**.39**
Word reading time	–.11	**–.58**	**–.55**	**–.40**	.10	.03	.09	–.21	**–.55**	**–.38**
Text reading errors	.12	**.42**	**.59**	**.28**	–.10	–.19	–.23	**.24**	**.50**	**.37**
Text reading time	–.14	**–.63**	**–.58**	**–.44**	.08	.04	.10	**–.31**	**–.59**	**–.42**
Text comprehension score	**–.29**	**–.26**	**–.40**	–.20	.12	.19	.22	–.23	**–.41**	**–.29**
Word spelling errors	.15	**.45**	**.67**	**.25**	–.11	–.10	–.20	**.31**	**.66**	**.41**
Text spelling errors	.13	**.49**	**.69**	**.33**	–.11	–.06	–.15	**.32**	**.70**	**.43**
RSPM raw score	.35	.08	.19	–.02	–.18	**–.26**	**–.33**	**.29**	**.28**	**.29**
WISC-III digit span raw score	–.08	**–.29**	**–.28**	–.16	.15	.18	**.24**	**–.27**	**–.34**	**–.34**
WISC-III arithmetic raw score	–.23	**–.27**	**–.38**	–.16	.14	.16	**.26**	–.16	**–.40**	**–.34**
Phoneme deletion errors	.06	.20	**.38**	.10	–.16	–.22	**–.26**	**.31**	**.39**	**.39**
Speech discrimination errors	.14	.22	**.32**	.08	**–.32**	**–.24**	**–.32**	**.36**	**.31**	.21

Note. RSPM = *Raven's Standard Progressive Matrices* (Raven, 1976); WISC-III = *Wechsler Intelligence Scale for Children,* 3rd ed., Greek version (WISC-III; Georgas et al., 1997). Computer-based assessment measures: TCS = text comprehension score; TRT = text reading time; TSE = text spell-checking errors; TST = text spell-checking time; FD = frequency discrimination; 2TS = two-tone sequencing; 3TS = three-tone sequencing; PD = pseudoword dictation; WPM = word–picture matching; LS = letter span.