Reading Research Quarterly
Vol. 42, No. 4
October/November/December 2007
© 2007 International Reading Association
(pp. 546–567)
doi:10.1598/RRQ.42.4.5

The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students

BRANT W. RIEDEL
Memphis City Schools Department of Research, Evaluation,
and Assessment, Tennessee, USA

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002) has gained widespread use in the United States as a measure of early reading skills. DIBELS has subtests designed to measure reading skills emphasized in the National Reading Panel report (National Institute of Child Health and Human Development, 2000) including phonemic awareness, phonics, fluency, and, to some degree, comprehension.

The creators of DIBELS assert that its subtests are useful for predicting future reading difficulty and facilitating early and accurate identification of students in need of intervention (Good, Simmons, & Kame'enui, 2001). DIBELS subtests intended to measure lower level reading skills such as phonological awareness (Phoneme Segmentation Fluency subtest, or PSF) and alphabetic principal (Nonsense Word Fluency subtest, or NWF) are administered in kindergarten and first grade to identify students at risk for reading difficulty and in need of intervention. Beginning in the middle of first grade, an additional subtest measuring students' speed and accuracy in reading connected text (Oral Reading Fluency subtest, or ORF) is administered to identify students in need of intervention.

Although DIBELS is being used in over 13,000 schools in the United States (according to the DIBELS website, https://dibels.uoregon.edu/data/index.php), often as part of the Reading First initiative, there is considerable controversy regarding the utility of the instrument. DIBELS's developers argue that the widespread use of DIBELS is supported by research, but its critics have suggested that political

ABSTRACTS

The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students

THE RELATION between Dynamic Indicators of Basic Early Literacy Skills (DIBELS) and reading comprehension at the end of first grade and second grade was examined in a sample of 1,518 first-grade students from a large urban school district. Receiver Operating Characteristic (ROC) analyses were used to determine optimal DIBELS cut scores for predicting satisfactory reading comprehension. A measure of reading rate and accuracy, a subtest that the DIBELS assessment refers to as Oral Reading Fluency (ORF), was a better predictor of comprehension than the remaining subtests, including a retell fluency task designed to measure comprehension. Also, use of other subtests in combination with ORF did not substantially improve predictive power beyond that provided by ORF alone. Vocabulary was an important factor in the relation between ORF scores and comprehension. Students with satisfactory ORF scores but poor comprehension had lower vocabulary scores than students with satisfactory ORF scores and satisfactory comprehension.

La relación entre DIBELS, comprensión lectora y vocabulario en estudiantes de primer grado en contextos urbanos

SE EXAMINÓ la relación entre los Indicadores Dinámicos de Habilidades Tempranas de Alfabetización (DIBELS) y la comprensión lectora al final de primero y segundo grado en una muestra de 1.518 estudiantes de primer grado de un gran distrito escolar urbano. Se usaron análisis ROC para determinar valores de corte de DIBELS óptimos para predecir la comprensión lectora satisfactoria. La medida de la velocidad y precisión en lectura, un subtest al que DIBELS denomina Fluidez de Lectura Oral (ORF), fue mejor predictor de la comprensión que los subtests restantes que incluyen una tarea de fluidez en el relato diseñada para medir comprensión. Asimismo el uso de otros subtests en combinación con ORF no mejoró sustancialmente el poder predictivo proporcionado por ORF solo. El vocabulario fue un factor importante en la relación entre los valores de ORF y comprensión. Los estudiantes con valores satisfactorios de ORF pero baja comprensión tuvieron valores de vocabulario más bajos que aquellos con valores satisfactorios de ORF y comprensión satisfactoria.

Die Relation zwischen DIBELS, dem Leseverständnis und Vokabular bei städtischen Schülern der ersten Klasse

DIE RELATION zwischen dynamischen Indikatoren der frühen Grundschreib- und Lesekenntnisse (DIBELS) und dem Leseverständnis am Ende der ersten und zweiten Schulklassen wurde an einem Beispiel von 1518 Schülern der ersten Klassen in einem großen städtischen Schulbezirk überprüft. Es wurden Kurvenanalysen Receiver Operating Charakteristic (ROC) verwendet, um optimale DIBELS Schnittstellenbenotungen zur Vorhersage eines befriedigenden Leseverständnisses zu ermitteln. Eine Bemessung von Lesebewertungen und deren Genauigkeit durch einen untergeordneten Test, welches in der DIBELS Bewertung als fließend mndliches Lesen oder Oral Reading Fluency (ORF) bezeichnet wird, stellte sich als besserer Verstndnisindikator heraus als die übrigen untergeordneten Tests, einschliesslich einer Aufgabe zur flssig vorgetragenen Wiederholung - konzipiert als Verständnisbemessung. Auch die Nutzung anderer Nebentests in Kombination mit ORF trugen nicht wesentlich zur Leistungssteigerung der Prognosequalitt bei, ausser bei jenen, die bereits durch ORF allein bestehen. Das Vokabular war ein wichtiger Faktor in der Relation zwischen ORF Benotungen und dem Verstehen. Schler mit zufriedenstellendem ORF aber schwachem Erfassen hatten niedrigere Vokabularwerte als Schler mit befriedigenden ORF Benotungen und befriedigendem Verständnis.

都市に住む一年生達のDIBELS、読解、語彙の関係

一学年と二学年の終わりの基礎的早期リテラシースキルの動的指標(DIBELS)と読解の関係が都市にある大規模の学区の一年生1,518人のサンプルで調査された。十分な読解を予測するために最適なDIBELS基準値を決定するために受信者操作特性(ROC)分析が使用された。読みの速度と正確さは、DIBELS評価がオーラル・リーディング・フルーエンシー(ORF)と呼んでいるサブテストで、理解を測定するように設計されたリテル・フルーエンシー・タスクを含む残りのサブテストよりも優れた予測を提供するものであった。また、ORFと合わせて行う他のサブテストの使用は、ORFのみによって提供されたもの以上には予測能力を向上しなかった。語彙は、ORFの点数と理解との関係において重要な因子であった。ORF得点は十分だが理解度が低い生徒達は、ORFの得点が高く理解度も高い生徒達より語彙力が低かった。

Les relations entre DIBELS, compréhension de la lecture et vocabulaire chez des élèves de première année

ON A EXAMINÉ dans un échantillon de 1518 élèves de première année d'une grande circonscription scolaire urbaine les relations entre les Indicateurs dynamiques des premières compétences de base en lettrisme (DIBEL) et la compréhension de la lecture en fin de première et de seconde année. L'analyse en Caractéristiques opératoires des récepteurs (ROC) a été employée pour définir les scores optimaux de coupure de DIBEL qui permettent une prédiction satisfaisante de compréhension de la lecture. L'évaluation de la vitesse et de la précision de la lecture à l'aide d'un sous-test que l'évaluation des DIBELS réfère comme mesure de Lecture courante à haute voix (ORF) s'avère être un meilleur prédicteur de la compréhension que les autres sous-tests, y inclus une tâche de fluidité du rappel utilisée comme mesure de la compréhension. De plus, l'utilisation des autres sous-tests en combinaison avec l'ORF n'améliore pas de façon substantielle le pouvoir prédictif au-delà de ce qu'apporte le seul ORF. Le vocabulaire est un facteur important de la relation entre les scores d'ORF et la compréhension. Les élèves ayant des scores d'ORF satisfaisants mais une compréhension limitée ont des scores de vocabulaire plus bas que les élèves ayant des scores d'ORF satisfaisants et une compréhension satisfaisante.

Связь между DIBELS, пониманием прочитанного и словарным запасом у первоклассников из городских школ

На основе выборки, состоявшей из 1518 первоклассников из бедных городских кварталов, исследовалось, как соотносятся *Динамические индикаторы основных первичных навыков грамотноасти* (DIBELS) и понимание прочитанного в конце первого и второго классов. При определении значений DIBELS, которые обеспечивают наилучшее понимание прочитанного, использовалась Характеристическая кривая (ROC-кривая). Скорость и точность чтения – особый тест, который, согласно DIBELS, именуется *беглостью чтения вслух* (ORF) – оказался более точным индикатором понимания прочитанного, чем все остальные тесты, включая задание на беглый пересказ текста, который был разработан непосредственно для проверки понимания. Кроме того, использование других тестов в сочетании с ORF не внесло существенных уточнений в прогноз, который был сделан на основе одного лишь ORF. Словарный запас оказался важным параметром для соотнесения результатов ORF с пониманием прочитанного. Учащиеся с удовлетворительным результатом ORF, но слабым пониманием прочитанного имели меньший словарный запас, чем учащиеся, продемонстрировавшие удовлетворительные результаты ORF и удовлетворительное понимание текста.

pressure to use DIBELS as part of Reading First is the reason for its widespread adoption (Goodman, 2006; Manzo, 2005).

One of the more common criticisms of DIBELS is that it is not an adequate indicator of reading comprehension (Goodman, 2006; Manzo, 2005). This criticism is important because both proponents and critics of DIBELS agree that comprehension is the ultimate goal of reading (Good et al., 2001; Goodman). If DIBELS subtests are closely connected to comprehension, they can be used to identify students at risk for comprehension difficulties and to provide additional instructional support to these students. If DIBELS subtests are not closely related to comprehension, misallocation of resources will occur. For example, students with good comprehension skills but low DIBELS scores will receive unnecessary intervention services, whereas students with high DIBELS scores but poor comprehension could be excluded from useful intervention.

It is not clear how closely reading comprehension is related to DIBELS tasks such as reading nonsense words (NWF) or pronouncing individual phonemes within words (PSF). Goodman (2006) provided a number of criticisms of these two DIBELS subtests. First, he disagreed with a stepping-stone model that suggests that certain reading skills, such as phoneme segmentation, must be mastered before moving to the next skills (e.g., fluency, comprehension). A related concern is that poor student performance on these subtests will lead to reading instruction being focused on these specific skills (phoneme segmentation, decoding nonsense words) at the expense of other instructional strategies that would help overall reading ability (Goodman; Pearson, 2006). Goodman also noted that dialect or articulation differences across teachers and students may make it difficult to consistently administer and score the NWF and PSF. Dialect issues may affect English-language learners (ELLs) to an even greater degree. In addition, more accomplished readers may be slowed on the timed test by attempts to make meaning of nonsense words and may be penalized for a tendency to say real words that are spelled similarly to the nonsense words (Goodman).

On the DIBELS ORF subtest, students read passages, and fluency is defined as the number of words read correctly in one minute. A student could read the words in the passage quickly, resulting in a high score, but still not comprehend the meaning of the text. Critics propose that the ORF task emphasizes speed rather than comprehension and may actually penalize students who are carefully searching for meaning within the text (Goodman, 2006; Pressley, Hilden, & Shankland, 2005). Samuels (2006) argued that ORF is not truly a measure of fluency because fluency involves decoding and comprehending at the same time, whereas the ORF task focuses on decoding speed but does not adequately assess comprehension. For example, ELLs may be able to decode text rapidly without comprehending the passage because of vocabulary difficulties (Samuels).

The DIBELS ORF task is followed by a retell fluency task designed to prevent students from speed reading without attempting to comprehend the passage. However, concerns have been raised about the ability to reliably score the retell fluency task and about its validity as a comprehension measure (Pressley et al., 2005).

An underlying concern that cuts across all DIBELS subtests is whether the information provided by DIBELS justifies the instructional time sacrificed to administer them. If DIBELS is a valid indicator of current and future reading comprehension ability, then its use could be justified for the purpose of screening, progress monitoring, and outcome assessment.

A number of theoretical and practical concerns regarding DIBELS have been raised in the preceding paragraphs. The widespread use of DIBELS for measuring progress and guiding instructional decisions makes it imperative for researchers to continue to examine the validity of the instrument. In the following paragraphs, empirical evidence supporting the use of DIBELS is examined, and gaps in the literature that need to be filled are identified.

The strongest empirical support exists for the DIBELS ORF subtest. There is a rich literature examining the use of an oral reading fluency task as a component of curriculum-based measurement (CBM; Fuchs, Fuchs, Hosp, & Jenkins, 2001). Both CBM ORF and DIBELS ORF involve reading connected text, and both operationally define fluency as the number of words read correctly in one minute. Although CBM and DIBELS ORF do not directly measure comprehension, results from multiple studies indicate that oral reading fluency, defined as number of words from connected text read correctly per minute, is significantly correlated with comprehension scores (Fuchs, Fuchs, & Maxwell, 1988; Fuchs et al., 2001). Two studies found significant correlations of .67 (Good et al., 2001) and .70 (Buck & Torgesen, 2003) between CBM ORF and state-mandated reading assessment scores for third-grade students. Also, support for an oral reading speed and comprehension relation is found outside of the CBM literature. A study using National Assessment of Educational Progress data from fourth-grade students found a positive relation between oral reading speed and reading comprehension (Daane, Campbell, Grigg, Goodman, & Oranje, 2005).

Findings with other measures of oral reading fluency may or may not generalize to DIBELS ORF, and therefore the specific passages and methods used as part of DIBELS ORF need to be examined directly. There is evidence that DIBELS ORF scores are significantly correlated with comprehension skills, at least among third-grade students. Technical reports have documented statistically significant correlations (ranging from .73 to .80) between third-grade students' scores on the DIBELS ORF and state-mandated assessments of reading (Barger, 2003; Shaw & Shaw, 2002; Wilson, 2005).

In contrast, Pressley and colleagues (2005) found a weaker correlation ($r = .45$) between DIBELS ORF and TerraNova Reading scores among third-grade students. Pressley et al. proposed that their weaker correlations may have occurred because the TerraNova is a more comprehensive test of reading achievement than state-mandated tests, which may focus on lower level reading skills. Consequently, Pressley et al. called for further studies of DIBELS ORF's relation with various reading measures outside of state-mandated tests.

The emphasis on studying third-grade students leaves open the question of how appropriate DIBELS ORF is for lower grades. Although it is recommended that administration of DIBELS ORF begin in first grade, I found only one study that examined the relation between first-grade ORF scores and reading comprehension (Cook, 2003). A second study of first-grade students examined an alternative form of ORF that also was developed by the creators of DIBELS (Roberts, Good, & Corcoran, 2005).

Among a sample of first-grade students ($n = 79$) in rural Ohio, Cook (2003) found a correlation of .73 between DIBELS ORF and the Stanford Achievement Test (9th edition) Reading Comprehension Cluster. Cook acknowledged the need for more studies on this topic given that her sample was relatively homogeneous with regard to socioeconomic status and included no minority students. Roberts et al. (2005) examined an alternative form of DIBELS ORF (VIP ORF) developed for the Voyager Universal Literacy Program. In a sample of 86 first-grade students drawn from an urban school system, VIP ORF scores were correlated at a statistically significant level ($r = .76$) with the Woodcock-Johnson Broad Reading Cluster, which includes letter–word identification tasks in addition to comprehension tasks. Both the Cook and the Roberts et al. studies examined concurrent relationships between first-grade ORF scores and comprehension. Therefore, the ability of first-grade ORF scores to predict future reading comprehension has not been established.

Few studies have investigated the relation between reading comprehension and the DIBELS phonological awareness (PSF) and alphabetic principle (NWF) tasks. The developers of DIBELS derived benchmarks for the PSF and NWF through extensive analyses looking at their relation to future DIBELS measures (e.g., ORF) (Good et al., 2001; Good, Simmons, Kame'enui, Kaminski, & Wallin, 2002). These are important analyses, but they do not directly address the question of whether PSF and NWF can predict reading comprehension versus fluency as defined by DIBELS.

Cook (2003) did find a statistically significant correlation between the Stanford Comprehension Cluster and both the PSF ($r = .38$) and NWF ($r = .61$) in first-grade students. An additional study found statistically significant correlations between the Woodcock-Johnson Total Reading Cluster and the PSF and NWF (Good et al., 2004). However, it should be noted that the Woodcock-Johnson Total reading cluster includes components other than comprehension, such as letter–word identification and reading fluency, which may have contributed to the significant correlations. Other researchers reported that the correlation between a word-identification fluency task using real words and comprehension was stronger than the correlation between DIBELS NWF and comprehension in a sample of at-risk first-grade students (Fuchs, Fuchs, & Compton, 2004). Two studies found no significant relation between first-grade PSF scores and the Stanford Diagnostic Reading Test (Johnson, 1996; Kaminski & Good, 1996).

The studies that have examined the relation between PSF, NWF, and other reading measures provide preliminary evidence of a relation between these phonological processing tasks and comprehension. However, the results for PSF are mixed, and generalizing from existing PSF and NWF results raises concerns. The samples were collected from rural or urban fringe schools with low minority populations (0–9%) and relatively low poverty rates (12–57% free and reduced-cost lunch) or focused exclusively on students identified as at risk for reading difficulties (Fuchs et al., 2004). Studies of students with a broad range of reading abilities in high poverty urban settings are needed.

Although the DIBELS Retell Fluency (RF) subtest is intended to assess comprehension, studies examining the relation between DIBELS RF and comprehension are sparse. Three studies have investigated the relation between DIBELS RF and a reading comprehension measure (McKenna & Good, 2003; Pressley et al., 2005; Roberts et al., 2005). Two studies found a statistically significant relation between RF and the comprehension measure, but in both studies the relation between RF and comprehension was substantially weaker than the correlation between DIBELS ORF and comprehension (McKenna & Good; Roberts et al.). Pressley et al. found no statistically significant relation between RF and comprehension and provided empirical evidence that it is difficult to reliably score the RF subtest. Further exploration of the usefulness of administering RF in addition to ORF is warranted.

In addition, previous studies investigating the relation between first-grade DIBELS measures and comprehension have not provided DIBELS cut scores that could be used to identify students likely to have current or future reading comprehension difficulties. The published benchmark scores for first grade are generally based on the relation between DIBELS subtests and future DIBELS subtest scores rather than DIBELS subtests' relation to comprehension (Good et al., 2001, 2002). It would be interesting to determine how comprehension-based cut scores compare with currently used benchmarks.

Also, no study identified the characteristics of students for whom DIBELS was a poor predictor of comprehension. For example, students with satisfactory ORF scores but low comprehension may have poor vocabulary skills, a literacy component not directly assessed by DIBELS. Such information would be useful to teachers as they attempt to determine the predictive validity of DIBELS results for particular students.

The current study addresses the gaps in the literature by examining assessment results for a sample of first-grade students (*n* = 1,518) from a large urban school district in the United States. Students in the sample were administered the DIBELS tests at the beginning, middle, and end of their first-grade year and also were assessed with a measure of reading comprehension at the end of first grade (GRA+DE; Williams, 2001) and second grade (TerraNova Reading; CTB/McGraw-Hill, 2003).

Specific research questions addressed include the following:

- How well do DIBELS subtests (e.g., PSF, NWF, ORF, RF) administered at the beginning, middle, and end of first grade predict reading comprehension at the end of first grade and end of second grade?
- Do the various DIBELS subtests differ in their ability to predict comprehension? Given concerns about the amount instructional time consumed by DIBELS, could certain DIBELS subtests be eliminated from the assessment protocol without reducing predictive power?
- What are the optimal DIBELS cut scores to use when attempting to predict if a student's reading comprehension will be satisfactory by the end of first grade and end of second grade?
- How do DIBELS cut scores derived from the current sample of urban first-grade students compare with published DIBELS benchmarks derived from other samples?
- What are the characteristics of students for whom DIBELS is a poor predictor of reading comprehension? Specifically, do demographic characteristics or other variables such as vocabulary skills distinguish these students?

Method

Participants

Participants were first-grade students in the Memphis City Schools district during the 2003–2004 school year. All students attended a school with a Reading Excellence Act (REA) grant and participated in REA-related assessments (DIBELS and GRA+DE). For the current study, students receiving special-education services were not included in the sample.

A total of 1,518 students were included in the sample. Students were predominately African American (*n* = 1,395, 92%), and the sample contained a nearly equal representation of females (*n* = 760) and males (*n* = 758). The poverty rate in the sample was high, with 85% of the students qualifying for free or reduced-cost lunch.

The population of ELLs within the participating schools was small, and therefore only a few ELL students were included in the sample (*n* = 59). All ELL students spoke Spanish as their primary language, but DIBELS and GRA+DE scores indicate that English-reading ability varied substantially across these students. Because of their small number and the reading challenges that are unique to this subgroup, results for ELL students were analyzed separately. For most of the analyses in this article, ELL students were excluded, but Pearson correlations between DIBELS and comprehension measures were calculated for ELL and non-ELL students, and the strength of the correlations was compared across the two subgroups.

Measures

DIBELS

The DIBELS Letter Naming Fluency (LNF), Phoneme Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), Oral Reading Fluency (ORF), and Retell Fluency (RF) subtests were administered.

In the LNF subtest, students are shown an 8.5" × 11" sheet of paper with randomly arranged upper- and lowercase letters. Students are asked to name as many letters as they can, and the LNF score is the number of letters correctly named in one minute.

For the PSF subtest, the test administrator orally presents words consisting of three to four phonemes. The student is prompted to say the individual phonemes for each word, and after each student response, the administrator says the next word. The number of points received for a particular word is the number of individual phonemes correctly produced by the student. The total PSF score is equal to the number of phonemes correctly produced by the student in one minute.

In the NWF subtest, students are shown an 8.5" × 11" sheet of paper with consonant–vowel–consonant and vowel–consonant "nonsense" words (e.g., *paj* or *ut*). The student is prompted to pronounce each nonsense word, and a point is given for each letter sound correctly produced by the student. The total NWF score is the number of letter sounds correctly produced in one minute.

For the ORF subtest, students are presented with a grade-level passage to read aloud. Specifically, the administrator says to the student, "Please read this out loud. If you get stuck, I will tell you the word so you can keep reading. When I say 'stop' I may ask you to tell me about what you read, so do your best reading. Start here. Begin." Students are given one minute to complete the task. This procedure is followed two more times with two additional grade-level passages. For each of the three passages, the score is the number of words read correctly in one minute. The median score from the three passages is used as the ORF score.

After each ORF passage was completed, the RF subtest was administered. For the RF task, students are asked to "Please tell me all about what you just read. Try to tell me everything you can. Begin." Then students are given one minute to tell the administrator about the passage. The administrator counts the number of words the student retells that illustrate an understanding of the passage, and this is the RF score for that passage. The overall RF score is the median RF score from the three passages.

Group Reading Assessment and Diagnostic Evaluation (GRA+DE)

The GRA+DE is a standardized, group-administered test of overall reading ability. The first-grade version of GRA+DE assesses vocabulary, comprehension, and oral-language skills, and questions are presented in a multiple-choice format. The total test score is derived from the vocabulary and comprehension components, and the oral-language (listening comprehension) section is optional. The vocabulary subtest requires students to select a word read by a teacher from a group of written words (word reading) and select a picture that matches a written word (word meaning). Thus, the vocabulary subtest requires reading and is not a measure of oral vocabulary. The comprehension subtest includes a sentence-comprehension component (select the word that best completes the sentence) and a passage-comprehension component. The GRA+DE is intended to be a test of power, not speed, and thus is not timed (Williams, 2001), but the assessment is typically completed in less than 90 minutes.

The publishers, American Guidance Service (AGS), report good psychometric properties for the instrument in their technical manual (Williams, 2001). The internal reliability coefficients (alpha) for the first-grade total test, vocabulary subtest, and comprehension subtest are .96, .93, and .94 respectively. Test–retest reliability for the total test at the first grade level is .96.

AGS reports high concurrent validity of the total test with other established instruments, such as the California Achievement Test at the first-grade level (.82) and the Gates-MacGinitie Reading Tests in a sample of first- and second-grade students (.90) (Williams, 2001). In addition to establishing validity for the total test, the technical manual reports good concurrent validity for the vocabulary and comprehension subtests, but these analyses focused on students in fifth grade or higher. Therefore, to examine concurrent validity of subtests in lower grades, we used a subsample of students from the present study who had taken both the GRA+DE and TerraNova in the spring of first grade ($n = 525$) and a sample of students from the same district who had taken the GRA+DE and TerraNova in the spring of second grade ($n = 1{,}418$). GRA+DE Comprehension subtest scores correlated at statistically significant levels with TerraNova Reading scores at the first- ($r = .63$) and second-grade levels ($r = .72$). In addition, GRA+DE Vocabulary subtests scores correlated at a statistically significant level with TerraNova Vocabulary scores in the spring of first grade ($r = .64$).

Fall and spring norms are available and are based on a diverse sample of over 16,000 students tested during the spring and fall of 2000. For the current study, the majority of the analyses focused on the comprehension subtest, and some analyses incorporated vocabulary subtest results.

TerraNova Reading subtest

The TerraNova is a widely used standardized test of achievement with timed subtests. The assessment is administered to groups and presents questions in a multiple-choice format. The Reading subtest of the TerraNova (second edition) was used as a measure of second-grade reading comprehension for the present study. The technical manual discusses the construct validity of the Reading subtest and reports a KR20 reliability coefficient of .89 for this subtest (CTB/McGraw-Hill, 2003). Norms for the TerraNova were derived from standardization with a diverse sample of more than 100,000 students tested in the fall of 1999 and spring of 2000. The Institute for the Development of Educational Achievement (IDEA) analyzed a number of K–3 reading assessments and concluded that there was sufficient evidence for using the TerraNova Reading subtest as a measure of reading comprehension in grades 1 through 3 (Kame'enui, 2002).

Procedure

In August of 2003, teachers received training in the administration of the DIBELS and GRA+DE. The district purchased DIBELS materials, which are published by Sopris West, from Pearson Scott Foresman, and representatives from Pearson provided DIBELS training for district teachers. AGS representatives provided teachers with training on the administration of GRA+DE. Administration of the TerraNova during the spring of second grade is mandated by the district, and school representatives attend training sessions provided by the district shortly before administration.

DIBELS and GRA+DE were administered during the 2003–2004 school year at the 26 district schools that had received a Reading Excellence Act grant. A district schedule was developed that defined the time period during which the DIBELS beginning-of-the-year (September 1–15, 2003), middle-of-the-year (January 2–16, 2004), and end-of-the-year (May 1–15, 2004) assessments were administered. Particular subtests were administered at certain times of the year as recommended by DIBELS's developers. Specifically, the LNF, PSF, and NWF were administered at the beginning of first grade and the PSF, NWF, ORF, and RF were administered at the middle and end of first grade.

Teachers were required to administer the GRA+DE at some point from April 15–30, 2004. TerraNova results were used to measure longer term (second grade) outcome for the current study and were based on the TerraNova assessment that occurred during the spring of 2005 (April 18–22).

Different assessments were used to measure comprehension in the first (GRA+DE) and second (TerraNova) grades for the following reasons. First, administration of the first-grade TerraNova is not required by the district, and therefore a large number of students in the current sample have no first-grade TerraNova results, but all have first-grade GRA+DE results. Second, the GRA+DE was not administered during the 2004–2005 school year, requiring that a different measure (TerraNova) be used to investigate second-grade outcomes.

Analysis approach

Receiver Operating Characteristic (ROC) analysis (Swets, Dawes, & Monahan, 2000) was used as the main approach for examining the relation between DIBELS subtests and reading comprehension. Through ROC analyses, each DIBELS measure available at each time period (beginning, middle, and end of the year) was examined as a predictor of reading comprehension status at the end of first grade.

ROC analysis was originally associated with electronic signal-detection theory, the study of an individual's accuracy in distinguishing various electronic signals. More recently, ROC has become widely used in psychology and medicine. For example, ROC has been employed to assess the accuracy of diagnostic tests and procedures in predicting breast and prostate cancer (Swets et al., 2000). The sensitivity and specificity of an assessment are important concepts in ROC analysis, and the definitions of these terms are reviewed below.

When using an assessment such as DIBELS to predict a dichotomous outcome (satisfactory or poor comprehension), four results are possible. Before listing the four possibilities it is important to note that medical terminology is used such that the term *positive* represents the existence of a problem (presence of a disease or in this case poor reading comprehension). First, DIBELS indicates that there is a comprehension problem, and there truly is a problem (true positive). Second, DIBELS indicates a problem, and there is not a problem (false positive). Third, DIBELS indicates no problem, and there is no problem (true negative). Fourth, DIBELS indicates no problem, and there is a problem (false negative).

These outcomes define the sensitivity and specificity of the assessment. In the current study, sensitivity refers to the proportion of poor-comprehension cases correctly predicted by a DIBELS subtest (true positive/true positive+false negative), whereas specificity is the proportion of satisfactory-comprehension cases correctly predicted by a DIBELS subtest (true negative/true negative+false positive).

There is a trade-off between sensitivity and specificity; as sensitivity increases, specificity decreases, and vice versa. For example, one could achieve a perfect specificity of 1.0 by setting the DIBELS ORF cut score at 0. All students would score 0 or above and would be predicted to have satisfactory comprehension. Thus all students with satisfactory comprehension would be correctly predicted. However, sensitivity would be very poor, with DIBELS incorrectly predicting satisfactory comprehension in all students with poor comprehension. The opposite problem could be created by setting the cut score too high.

ROC curves are plotted by graphing sensitivity (y axis) against 1 – specificity (x axis), and one can identify cut scores on the assessment (e.g., DIBELS) that maximize sensitivity and specificity and can guide decision making. An ideal ROC curve, suggesting a highly predictive assessment, would move steeply up the y axis toward the upper left-hand corner, indicating there is a cut point where both sensitivity and specificity are high (and 1 – specificity is low). In contrast, a poor assessment would tend to produce a diagonal line from the bottom left-hand corner to the upper right-hand corner of the graph, indicating no cut point at which both sensitivity and specificity are high.

In determining cut scores for the current article, an attempt was made to strike a balance between sensitivity and specificity, with a slight bias toward sensitivity. Specifically, the cut score for each DIBELS subtest was defined as the score at which there was the smallest difference between sensitivity and specificity, and sensitivity was equal to or greater than specificity. For example, if one cut score provided a sensitivity of .69 and a specificity of .72, and the next cut score offered a sensitivity of .72 and a specificity of .69, the latter cut score was chosen.

It should be noted that in some situations either sensitivity or specificity may be more important, and one would select cut scores that emphasized sensitivity or specificity rather than balancing the two. In the current study, cut scores were chosen to balance the competing desires of identifying all students with comprehension difficulties and ensuring that limited intervention resources are directed only toward students with comprehension difficulties.

Based on GRA+DE comprehension subtest scores at the end of first grade, students were classified as being in the satisfactory-comprehension group (Normal Curve Equivalent or NCE $\geq$ 40; equivalent to a percentile of approximately 32 or above) or the poor-comprehension group (NCE < 40). GRA+DE results were not available during the following school year (2004–2005), but most of the students in the sample took the TerraNova Reading subtest at the end of second grade, which provided data for examining the relation between first-grade DIBELS scores and reading comprehension the following year. On the TerraNova Reading subtest, satisfactory comprehension was again defined as an NCE $\geq$ 40, and poor comprehension was defined as an NCE below 40.

Determining the definition of satisfactory comprehension is admittedly a subjective process. A criterion of 40 NCE or higher was chosen for the following reasons. First, a score of 40 NCE is within one half of a standard deviation of the median score for the national normative sample. Second, a cut point of 40 NCE provides a stricter definition of poor comprehension relative to other possible cut points such as the 40th percentile or a 50 NCE percentile. If the emphasis is on identifying those most at risk for comprehension difficulties, then the lower cut point (40 NCE) is preferable to higher cut points (e.g., 50 NCE).

One could also argue that a 40 NCE criterion for satisfactory comprehension is too liberal. Therefore, the ROC analyses were repeated using an alternative, higher cut point (40th percentile), and these findings are also reported in the Results section.

Several outcomes from the ROC analyses are reported including area under the curve (AUC) and percent of students classified correctly, both general measures of the predictive ability of the variable being examined (i.e., a DIBELS subtest). The AUC can range from 0.5 to 1.0, with 0.5 indicating accuracy no better than chance and 1.0 indicating perfect predictive accuracy (Swets et al., 2000). Sensitivity, specificity, and a cut score are also reported.

In addition to examining individual DIBELS subtests, the use of subtest combinations was also examined through logistic regression. Logistic regression is widely used in the medical literature, and like linear regression, it involves identifying significant predictors of a dependent variable. However, logistic regression is more appropriate than linear regression when the dependent variable is dichotomous (Menard, 1995). Performance on the GRA+DE comprehension subtest and TerraNova Reading subtest (NCE ≥ 40 or not) were used as the dependent variables for the logistic regression analyses, and DIBELS subtest scores served as predictor variables. A forward stepwise procedure was used to determine if the predictive power of DIBELS improved as subtests were added to the regression equation.

Also, ANOVA, chi-square, and logistic regression analyses were used to examine students for whom DIBELS was a poor predictor of comprehension. To facilitate comparisons with previous papers, Pearson correlations were calculated between DIBELS subtests and comprehension measures.

For all analyses at a particular time period, only students who had scores for every DIBELS measure collected at the time period were included in the analyses. For example, to be included in the beginning-of-the-year analyses, a student was required to have an LNF, PSF, and NWF score (plus a spring GRA+DE score). This allowed for a fair comparison of the predictive strength of different DIBELS subtests at a particular time period (e.g., middle-of-the-year NWF vs. middle-of-the-year ORF in predicting comprehension). SPSS version 14.0 was used for all statistical analyses.

Results

Individual DIBELS subtests and first-grade comprehension

The top section of Table 1 summarizes the results from the ROC analyses investigating the relation between beginning-of-the-year DIBELS scores and end-of-first-grade comprehension scores. The NWF score proved to be a slightly better predictor of comprehension than the PSF and LNF subtests. The middle section of Table 1 outlines the ROC results when middle-of-the-year DIBELS scores were examined. At this point students begin taking the ORF subtest, and it proved to be the best single predictor of comprehension at the end of first grade.

The final section of Table 1 shows ROC results for end-of-the-year DIBELS. ORF continues to be a good predictor of reading comprehension, classifying 80% of the students correctly. In contrast, the PSF score is a poor predictor of reading comprehension, classifying just over half of the students correctly.

Individual DIBELS subtests and second-grade comprehension

In addition to results within the first-grade year, longer term results were considered by using

TABLE 1
ROC RESULTS FOR BEGINNING-, MIDDLE-, AND END-OF-FIRST-GRADE DIBELS'S ABILITY TO PREDICT SATISFACTORY COMPREHENSION (GRA+DE NCE ≥ 40) AT THE END OF FIRST GRADE

	Area under curve	Cut score	Sensitivity	Specificity	Percentage classified correctly
Beginning-of-year DIBELS (n = 1,274)					
Beginning LNF	.73	37	.68	.65	66
Beginning PSF	.66	16	.61	.60	61
Beginning NWF	.74	16	.68	.68	68
Middle-of-year DIBELS (n = 1,027)					
Middle PSF	.59	31	.56	.56	56
Middle NWF	.73	32	.68	.65	66
Middle ORF	.83	18	.77	.76	77
Middle RF	.75	6	.69	.67	68
End-of-year DIBELS (n = 1,224)					
End PSF	.57	41	.55	.52	53
End NWF	.74	39	.70	.67	68
End ORF	.87	30	.80	.80	80
End RF	.77	14	.71	.68	69

end-of-second-grade comprehension scores. For the majority of the sample, TerraNova Reading scores were available from the spring of second grade. ROC results using the second-grade TerraNova are summarized in Table 2.

Similar to the first-grade comprehension results, beginning-of-first-grade LNF and NWF were slightly better than PSF at predicting second-grade comprehension. Also consistent with previous analyses, the middle-of-the-year and end-of-the-year first-grade ORF proved to be better predictors of second-grade comprehension than the remaining DIBELS subtests. Given the increased length of time between the predictor (ORF) and the event being predicted (comprehension), it was not surprising that the end-of-first-grade ORF was less successful at predicting second-grade comprehension than first-grade comprehension. Still, the percentage of students classified successfully was relatively high (71%).

DIBELS subtest combinations and comprehension

Logistic regression analyses were conducted to examine DIBELS subtest combinations. Forward stepwise regression was used, with the likelihood-ratio statistic employed as the test of significance. For both the first- and second-grade comprehension analyses, the dependent variable was dichotomously coded to indicate satisfactory (NCE $\geq$ 40) or poor (NCE < 40) comprehension. Results are summarized in Table 3.

For all of the middle- and end-of-year analyses, ORF was the first variable entered by the stepwise procedure because of the strength of its relation with comprehension. In each analysis a subtest other than ORF was statistically significant and therefore entered the equation in step 2, but in no analysis did a variable other than ORF produce a practically important increase in predictive accuracy. In all cases, the added subtest or subtests increased predictive accuracy by less than 1 percentage point.

Analyses repeated using percentile of 40 as an indicator of satisfactory comprehension

The ROC and logistic regression analyses were repeated using a percentile of 40 or higher as an alternative indicator of satisfactory comprehension. ROC results using the 40th percentile criterion are summarized in Tables 4 and 5. Not surprisingly, recommended DIBELS cut scores are slightly higher because the 40th percentile repre-

TABLE 2
ROC RESULTS FOR BEGINNING-, MIDDLE-, AND END-OF-FIRST-GRADE DIBELS'S ABILITY TO PREDICT SATISFACTORY COMPREHENSION (TERRANOVA READING NCE $\geq$ 40) AT THE END OF SECOND GRADE

	Area under curve	Cut score	Sensitivity	Specificity	Percentage classified correctly
Beginning-of-year DIBELS (*n* = 1,112)					
Beginning LNF	.70	39	.67	.64	65
Beginning PSF	.62	17	.58	.58	58
Beginning NWF	.68	19	.65	.61	63
Middle-of-year DIBELS (*n* = 891)					
Middle PSF	.58	32	.54	.54	54
Middle NWF	.65	34	.62	.58	60
Middle ORF	.76	22	.69	.65	67
Middle RF	.71	8	.70	.63	66
End-of-year DIBELS (*n* = 1,054)					
End PSF	.60	41	.59	.59	59
End NWF	.70	41	.67	.64	65
End ORF	.78	34	.71	.71	71
End RF	.73	16	.69	.65	66

TABLE 3
STEPWISE LOGISTIC REGRESSION WITH READING COMPREHENSION AS THE DEPENDENT VARIABLE AND DIBELS MEASURES AS PREDICTORS

	Beginning-of-first-grade DIBELS	Middle-of-first-grade DIBELS	End-of-first-grade DIBELS
DV = GRA+DE Comprehension (first grade)*			
Regression step 1: Significant predictors	NWF	ORF	ORF
Regression step 1: Classification accuracy	68.8 %	77.4%	79.5%
Regression step 2: Significant predictors	NWF, LNF	ORF, NWF	ORF, RF
Regression step 2: Classification accuracy	71.2 %	77.5%	79.7%
DV = TerraNova Reading (second grade)*			
Regression step 1: Significant predictors	LNF	ORF	ORF
Regression step 1: Classification accuracy	66.5%	67.9%	71.8%
Regression step 2: Significant predictors	LNF, NWF	ORF, RF	ORF, RF
Regression step 2: Classification accuracy	66.2%	68.7%	72.4%
Regression step 3: Significant predictors	—	—	ORF, RF, PSF
Regression step 3: Classification accuracy	—	—	72.4%

*Dependent variables were dummy coded as 1 for an NCE < 40 and 0 for an NCE ≥ 40.

TABLE 4
ROC RESULTS FOR BEGINNING-, MIDDLE-, AND END-OF-FIRST-GRADE DIBELS'S ABILITY TO PREDICT END OF FIRST-GRADE COMPREHENSION (GRA+DE COMPREHENSION PERCENTILE ≥ 40)

	Cut score	Percentage classified correctly
Beginning-of-year DIBELS (*n* = 1,274)		
Beginning LNF	39	65
Beginning PSF	18	61
Beginning NWF	18	68
Middle-of-year DIBELS (*n* = 1,027)		
Middle PSF	33	54
Middle NWF	33	67
Middle ORF	21	77
Middle RF	7	67
End-of-year DIBELS (*n* = 1,224)		
End PSF	41	53
End NWF	40	68
End ORF	34	79
End RF	16	69

TABLE 5
ROC RESULTS FOR BEGINNING-, MIDDLE-, AND END-OF-FIRST-GRADE DIBELS'S ABILITY TO PREDICT END OF SECOND-GRADE COMPREHENSION (TERRANOVA READING PERCENTILE ≥ 40)

	Cut score	Percentage classified correctly
Beginning-of-year DIBELS (*n* = 1,112)		
Beginning LNF	40	64
Beginning PSF	18	58
Beginning NWF	20	63
Middle-of-year DIBELS (*n* = 891)		
Middle PSF	33	56
Middle NWF	35	61
Middle ORF	23	67
Middle RF	9	66
End-of-year DIBELS (*n* = 1,054)		
End PSF	42	58
End NWF	42	65
End ORF	38	72
End RF	17	66

sents a higher level of reading comprehension than a 40 NCE. In general, DIBELS cut scores were 1 to 2 points higher when using the 40th percentile as the indicator of satisfactory comprehension, but in two analyses DIBELS cut scores were 4 points higher when the 40th percentile method was employed.

The general pattern of results was the same regardless of which definition of satisfactory comprehension was used. Beginning-of-the-year LNF and NWF were better predictors than PSF. For the middle- and end-of-the-year DIBELS, ORF again was the best predictor of first- and second-grade comprehension.

Results from logistic regressions examining subtest combinations were also similar to those found when using the 40 NCE criterion. Subtests entering the equation after ORF increased predictive accuracy by less than 1 percentage point, with one exception. When middle-of-the-year NWF was added to the equation containing middle-of-the-year ORF, accuracy in predicting first-grade comprehension improved by 1.9 percentage points (75.7% to 77.6%).

Correlations between DIBELS subtests and comprehension

To this point in this report the predictive ability of DIBELS subtests has been framed as the ability to predict a dichotomous outcome (poor versus satisfactory comprehension). The Pearson correlation analyses summarized in Table 6 treat comprehension as a continuous variable and provide an alternative method for analyzing the strength of the relation between DIBELS and reading comprehension. Only eight ELL students took the TerraNova in second grade, and therefore only first-grade results are reported for this subgroup. Results found using correlations resemble those found with ROC analyses. ORF is the subtest most strongly related to comprehension, and PSF has a weak relation with comprehension. Interestingly, the relation between ORF and comprehension was stronger for ELL students than for non-ELL students.

Characteristics of students misclassified by DIBELS ORF

Analyses were conducted to determine if there were statistically significant differences between students whose DIBELS end-of-first-grade ORF correctly predicted end-of-first-grade comprehension status and students whose end-of-the-year ORF resulted in a misclassification of comprehension status. Specifically, vocabulary, gender, and poverty status were investigated as characteristics that might be associated with misclassification of comprehension status. These analyses focused on concurrent results (spring ORF and spring comprehension) to eliminate the influence of time-related variables that could have weakened the ability of ORF to predict comprehension.

For vocabulary, a 2 (ORF status, satisfactory or poor) × 2 (comprehension status, satisfactory or poor) ANOVA was used to examine end-of-the-year GRA+DE Vocabulary subtest NCE scores. Based on the ROC analyses detailed previously, a satisfactory ORF score was defined as 30 or above, whereas a satisfactory comprehension score was defined as an NCE of 40 or above.

The main effects for ORF status, $F(1, 1220) = 135.8$, $p < .001$, and comprehension status, $F(1, 1220) = 271.7$, $p < .001$, were significant. The interaction effect also was significant, $F(1, 1220) = 14.1$, $p < .001$. All follow-up tests of simple effects were significant, $p < .001$. One striking finding was the

TABLE 6
PEARSON CORRELATIONS BETWEEN DIBELS SUBTESTS AND FIRST-GRADE GRA+DE COMPREHENSION AND SECOND-GRADE TERRANOVA READING

	GRA+DE (not ELL)	GRA+DE (ELL)	TerraNova (not ELL)
Beginning-of-year DIBELS			
Beginning LNF	.44	.15	.40
Beginning PSF	.26	.31	.26
Beginning NWF	.45	.41	.39
Middle-of-year DIBELS			
Middle PSF	.16	.19	.18
Middle NWF	.45	.47	.38
Middle ORF	.59	.72	.49
Middle RF	.41	.42	.39
End-of-year DIBELS			
End PSF	.15	.21	.23
End NWF	.46	.41	.37
End ORF	.67	.80	.54
End RF	.51	.69	.46

vocabulary difference between the two satisfactory ORF groups. Although both groups performed well on the ORF task, the group with poor comprehension scored over 20 NCE points lower on the vocabulary subtest, *M* NCE = 35.2, *n* = 100, than the group with satisfactory comprehension, *M* NCE = 57.5, *n* = 587. Similarly, those with poor ORF scores who still managed to do well on the comprehension test had significantly better vocabulary, *M* NCE = 40.5, *n* = 149, than those with poor ORF scores who did not fare well on the comprehension test, *M* NCE = 26.5, *n* = 388.

Chi-square analyses were used to examine the relation between the four possible ORF/comprehension outcomes and two other variables, gender and poverty status. The chi-square for gender was significant, $\chi^2(3) = 24.6$, $p < .001$, with the main finding being that students with satisfactory ORF scores but poor comprehension scores were more likely to be males. Among students with ORF scores of 30 or above, a significantly greater percentage of males (18%) than females (12%) scored below a 40 NCE on the comprehension test. The chi-square analysis for poverty status was not significant, $\chi^2(3) = 4.5$, $p = .21$.

Given that there was substantial variability in ORF scores in the satisfactory ORF category (scores of 30 and 90 would both count as satisfactory) and poor ORF category, additional analyses were conducted to see if vocabulary and gender remained statistically significant predictors of comprehension outcome when controlling for ORF scores.

A logistic regression was conducted with GRA+DE end-of-the-year comprehension NCE status (≥ 40 or < 40) as the dependent variable and end-of-the-year ORF, end-of-the-year GRA+DE vocabulary NCE, gender, and poverty status as the predictor variables. Forced entry was used to include ORF in the initial step of the analysis. In the second step of the analysis, vocabulary, gender, and poverty status were examined with forward stepwise regression using the likelihood-ratio statistic as the test of significance. ORF was a statistically significant predictor and by itself produced a classification accuracy rate of 79.5%. In the second step, vocabulary was the only predictor that met entry criteria, and it improved the classification accuracy rate to 82.7%.

Discussion

DIBELS ORF administered in first grade proved to be a good predictor of reading comprehension at the end of first grade and second grade. DIBELS ORF collected at the end of first grade was able to predict students' first- and second-grade reading comprehension status (satisfactory or not) with 80% and 71% accuracy, respectively. The remaining DIBELS subtests were less accurate predictors of comprehension, with PSF being the weakest predictor. PSF results collected at the middle and end of first grade predicted first- and second-grade comprehension at a rate that was only slightly better than chance. Other studies also have found relatively weak PSF correlations with comprehension (Johnson, 1996; Kaminski & Good, 1996).

In addition, considering other subtest results in combination with ORF results did not substantially improve on the predictive accuracy produced by ORF alone. If the goal of DIBELS administration is to identify students at risk for reading comprehension difficulties, the present results suggest that by the middle of first grade, administration of DIBELS subtests other than ORF is not necessary. The minimal gains in predictive accuracy do not justify the time and effort involved in administering the non-ORF subtests to each student.

The present study does not directly address the question of why ORF is more highly correlated with reading comprehension than are other DIBELS subtests. However, there are a number of plausible hypotheses. First, in comparison to identifying individual phonemes or reading isolated nonsense words, reading connected text with real words more closely mimics a typical reading comprehension task. Second, the ability to read connected text rapidly and accurately may play a crucial role in one's ability to comprehend text, resulting in a close relation between comprehension ability and reading rate measures such as ORF. Fuchs et al. (2001) discussed two prominent reading theories and pointed out that both postulate that rapid word recognition frees up cognitive resources for higher level comprehension processes. Third, concerns have been raised about the ability to reliably score the PSF and RF subtests (Goodman, 2006; Pressley et al., 2005). Inconsistent scoring across test administrators and across students with the same administrator would reduce the strength of the measures' correlations with a criterion variable such as comprehension. Fourth, by the middle of first grade, even students significantly behind their peers in reading development may have mastered the lower level skills required by the PSF, making the PSF less likely to distinguish between good and poor readers. Finally, Goodman proposed that good readers could be penalized on the NWF subtest if they spend time trying to make sense of the nonsense words or if their knowledge of real words makes them mispronounce nonsense words resembling actual words.

The correlation between DIBELS ORF and a comprehension task when both were administered in the spring of first grade was .67, slightly lower than the correlation (.73) found between DIBELS ORF and comprehension in the previous study of first-grade students (Cook, 2003). The correlation from the present study fell in the range of correlations reported for DIBELS or CBM ORF and state-mandated reading tests in third-grade students (.67–.80), and was substantially higher than the correlation (.45) in third-grade students reported by Pressley et al. (2005). Thus there is evidence that DIBELS ORF is able to predict comprehension ability in first grade as well as it does in third grade. It should be noted that the GRA+DE comprehension subtest is not a timed test, which may have increased the number of students in the present study who performed poorly on the timed ORF task but were able to perform well on the comprehension task. Overall this would tend to weaken the correlation between ORF and comprehension, resulting in a conservative test of their relationship.

Concerns have been expressed about using DIBELS ORF with English-language learners who may be able to decode words rapidly but do not comprehend text because of vocabulary problems. Thus, there are concerns that DIBELS scores could overestimate comprehension ability. However, the present study indicated that DIBELS ORF and comprehension were more strongly correlated in ELL students than in non-ELL students. Because of the small size of the ELL sample, further investigations are needed to determine if this result can be replicated.

Although the retell fluency (RF) task is intended to be an indicator of comprehension, RF was a weaker predictor of comprehension than ORF, and examining RF results in combination with ORF results did not substantially improve predictive accuracy over using ORF alone, findings that are consistent with previous studies (McKenna & Good, 2003; Pressley et al., 2005; Roberts et al., 2005). Nevertheless, including the RF task with ORF could strengthen the relation between ORF scores and comprehension because the RF task indicates that one is to read for understanding and not just speed. The present study cannot provide evidence regarding this hypothesis because all students participated in the RF task. However, a previous study suggested that inclusion of a retell cue in the ORF instructions does not substantially affect the relation between ORF scores and comprehension (McKenna & Good, 2003). Thus, there remains a lack of empirical evidence for the usefulness of the RF task.

A comparison of the cut scores derived from the present sample and published DIBELS benchmarks derived from other samples (Good et al., 2002) suggests that the cut scores and published benchmarks are generally in agreement. The published benchmarks divide students into three categories (at risk, some risk, low risk; or deficit, emerging, established, depending on the timing of the measure), whereas cut scores divide students into two categories. In most cases cut scores derived from the present sample fell in the middle category of the published benchmarks (either some risk or emerging).

However, there were some exceptions. In three of four LNF analyses the recommended cut score for LNF was higher than the published minimum requirement for being labeled low risk (a score of 37). Similarly, in three of four analyses the recommended middle-of-the-year ORF cut score was higher than the published low-risk requirement for ORF (a score of 20). Therefore, a slight increase in the low-risk requirements for LNF and middle-of-the-year ORF may be needed in urban school populations.

When deriving kindergarten and first-grade benchmarks, DIBELS's developers decided to use an ORF score of 40 as the benchmark for the end of first grade. Good et al. (2001) indicated that this end-of-first-grade benchmark was based on "empirical, theoretical, and social-validation sources" (p. 266). The selection of earlier DIBELS benchmarks (i.e., kindergarten and early first grade) was based on their ability to predict this end-of-first-grade benchmark ORF of 40. Thus, the selection of the 40 ORF benchmark was an important assumption in the process of developing all of the kindergarten and first-grade DIBELS benchmarks. The present study results suggest that an end-of-first-grade ORF of 40 would be associated with comprehension success at the end of second grade. If second-grade comprehension success is defined as scoring at the 40th percentile or higher, the present study suggests an ORF cut score of 38 at the end of first grade, which is similar to the published benchmark of 40.

One could argue the relative usefulness of the published benchmarks that divide students into three categories or the present cut scores that separate students into two groups. The published benchmarks are useful in that they help to identify students at a very high risk for reading failure and students at a very low risk for such failure. However, the middle category tends to be quite wide, leaving one with a large group of students for whom the necessity of intervention is not clear. Cut scores derived from the present sample attempt to define the optimal point at which one could divide students into those needing intervention or not. Of course the dichotomous decision of intervention or no intervention does not suggest that everyone below the cut score receives the same intervention. More intensive intervention could be offered to students falling farther below the cut point. In addition, if resources permit, one could choose to offer intervention to those scoring at or slightly above the cut point.

The present results also point to the important role of vocabulary in reading comprehension. Approximately 15% of students in the present sample with satisfactory ORF scores at the end of first grade had poor comprehension at the same time. A striking characteristic of the students with satisfactory ORF but poor comprehension was poor vocabulary skills, *M* NCE = 35.2, especially relative to those students with satisfactory ORF and comprehension, *M* NCE = 57.5. Thus, being aware of students' vocabulary abilities either through careful observation or more formal testing may help teachers when interpreting ORF scores. These findings also suggest that for some students, vocabulary needs to be the focus of intervention.

A limitation of the study should be noted. The participating schools were recipients of a Reading Excellence Act grant that encouraged using assessment to drive instruction, provided a significant amount of professional development for teachers, and funded interventions to address reading difficulties. Therefore, it is possible that instructional practices or interventions applied in response to students' poor DIBELS scores could have affected the predictive power of DIBELS. Specifically, students in the current setting who performed poorly on DIBELS may have been more likely to meet future comprehension goals than students in other settings. However, it should be noted that results found with concurrent analyses of DIBELS and comprehension (where this concern would not be relevant) were similar to results found for analyses using DIBELS to predict future comprehension.

In summary, DIBELS ORF scores collected in first grade were a good predictor of first- and second-grade reading comprehension, but other DIBELS subtests were less successful at predicting students' comprehension. The relatively strong relation between DIBELS ORF and comprehension supports the use of DIBELS ORF as a screening (middle of first grade) and outcome measure (end of first grade). The value of DIBELS ORF as a diagnostic assessment is less clear. Although DIBELS ORF usually correctly predicts current and future comprehension difficulties, it may not provide any details regarding the student's reading difficulties or the interventions needed to remedy them. Also, further study is needed to identify factors that cause DIBELS ORF to misjudge comprehension ability in some students.

In contrast to the positive results found for DIBELS ORF, results from the present study support critics' concerns about the PSF and NWF subtests (Goodman, 2006). Cut scores derived from PSF and NWF scores at the end of first grade misjudged respectively the concurrent comprehension status of 47% and 32% of the students. The current results do not support intervention instruction in phoneme segmentation or decoding for those who score poorly on the PSF or NWF. From the middle of first grade to the end of first grade, an abbreviated DIBELS protocol is recommended that includes ORF but excludes NWF and PSF. This revised protocol would minimize the amount of instructional time lost and still preserve predictive power.

BRANT W. RIEDEL is an evaluator in the Memphis City Schools Department of Research, Evaluation, and Assessment (2597 Avery Avenue, Memphis, TN 38112, USA; e-mail riedelb@mcsk12.net). He received his PhD in psychology from the University of Memphis. Current research interests include reading assessment, formative assessment, evaluation of school-based literacy and health initiatives, and randomized experiments in school settings.

REFERENCES

BARGER, J. (2003). *Comparing the DIBELS Oral Reading Fluency indicator and the North Carolina end of grade reading assessment* (Tech. Rep.). Asheville: North Carolina Teacher Academy.

BUCK, J., & TORGESEN, J. (2003). *The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test* (FCRR Tech. Rep. No. 1). Tallahassee: Florida Center for Reading Research. Available: www.fcrr.org/TechnicalReports/TechnicalReport1.pdf

COOK, R.G. (2003). *The utility of DIBELS as a curriculum based measurement in relation to reading proficiency on high-stakes tests.* Unpublished master's thesis, Marshall University Graduate College, Huntington, WV.

CTB/MCGRAW-HILL. (2003). *TerraNova second edition: California Achievement Tests technical report.* Monterey, CA: Author.

DAANE, M.C., CAMPBELL, J.R., GRIGG, W.S., GOODMAN, M.J., & ORANJE, A. (2005, October). *Fourth-grade students reading aloud: NAEP 2002 special study of oral reading.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Available: nces.ed.gov/nationsreportcard/pdf/studies/2006469.pdf

FUCHS, L.S., FUCHS, D., & COMPTON, D.L. (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children, 71*, 7–21.

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.

FUCHS, L.S., FUCHS, D., & MAXWELL, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*(2), 20–28.

GOOD, R.H., & KAMINSKI, R.A. (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the

Development of Educational Achievement. Available: http://dibels.uoregon.edu.

GOOD, R.H., KAMINSKI, R.A., SHINN, M., BRATTEN, J., SHINN, M., LAIMON, D., SMITH, S., & FLINDT, N. (2004). *Technical adequacy of DIBELS: Results of the Early Childhood Research Institute on measuring growth and development* (Tech. Rep. No. 7). Eugene: University of Oregon.

GOOD, R.H., SIMMONS, D.C., & KAME'ENUI, E.J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high stakes outcomes. *Scientific Studies of Reading*, *5*, 257–288.

GOOD, R.H., SIMMONS, D.C., KAME'ENUI, E.J., KAMINSKI, R.A., & WALLIN, J. (2002). *Summary of decision rules for intensive, strategic, and benchmark instructional recommendations in kindergarten through third grade* (Tech. Rep. No. 11). Eugene: University of Oregon.

GOODMAN, K.S. (2006). A critical review of DIBELS. In K.S. Goodman (Ed.), *The truth about DIBELS: What it is, what it does* (pp. 1–39). Portsmouth, NH: Heinemann.

JOHNSON, D.S. (1996). *Assessment for the prevention of early reading problems: Utility of Dynamic Indicators of Basic Early Literacy Skills for predicting future reading performance.* Unpublished doctoral dissertation, University of Oregon, Eugene.

KAME'ENUI, E.J. (2002). *An analysis of reading assessment instruments for K–3.* Eugene, OR: Institute for the Development of Educational Achievement.

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

MANZO, K.K. (2005, September 28). National clout of DIBELS test draws scrutiny: Critics say reading tool's scope fails to justify its broad use. *Education Week*, *25*, 1, 12.

MCKENNA, M.K., & GOOD, R.H. (2003). *Assessing reading comprehension: The relation between DIBELS Oral Reading Fluency, DIBELS Retell Fluency, and Oregon State Assessment scores* (Tech. Rep.). Eugene: University of Oregon.

MENARD, S. (1995). *Applied logistic regression analysis.* Thousand Oaks, CA: Sage.

NATIONAL INSTITUTE OF CHILD HEALTH AND HUMAN DEVELOPMENT. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

PEARSON, P.D. (2006). Foreword. In K.S. Goodman (Ed.), *The truth about DIBELS: What it is, what it does* (pp. v–xix). Portsmouth, NH: Heinemann.

PRESSLEY, M., HILDEN, K., & SHANKLAND, R. (2005). *An evaluation of end-grade-3 Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Speed reading without comprehension, predicting little* (Tech. Rep.). East Lansing, MI: Michigan State University, Literacy Achievement Research Center.

ROBERTS, G., GOOD, R., & CORCORAN, S. (2005). Story retell: A fluency-based indicator of reading comprehension. *School Psychology Quarterly*, *20*, 304–317.

SAMUELS, S.J. (2006, May). *Introduction to reading fluency.* Paper presented at the annual meeting of the International Reading Association, Chicago.

SHAW, R., & SHAW, D. (2002). *DIBELS Oral Reading Fluency–based indicators of third grade reading skills for Colorado State Assessment Program (CSAP)* (Tech. Rep.). Eugene: University of Oregon.

SWETS, J.A., DAWES, R.M., & MONAHAN, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26.

WILLIAMS, K.T. (2001). *Technical manual: Group Reading Assessment and Diagnostic Evaluation.* Circle Pines, MN: American Guidance Service.

WILSON, J. (2005). *The relationship of Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Oral Reading Fluency to performance on Arizona Instrument to Measure Standards (AIMS)* (Tech. Rep.). Tempe, AZ: Tempe School District No. 3.

AUTHOR'S NOTE

This research was supported by Reading Excellence Act grants awarded to the participating schools. Appreciation is extended to the teachers and literacy leaders who assisted with administration of the assessments used in the study and to Henrietta Boykins, Susan Dold, and Linda Kennard, who coordinate elementary literacy interventions and assessments for the district. Appreciation is also extended to evaluators in the MCS Office of Research and Evaluation who provided feedback on earlier versions of the article.

Received June 8, 2006
Final revision received April 16, 2007
Accepted March 19, 2007