FOREWORD BY **P. DAVID PEARSON**

The Truth About

DIBELS

What It Is

What It Does

KENNETH S. GOODMAN

with

Alan Flurkey | Tsuguhiko Kato | Constance Kamii

Maryann Manning | Susan Seay | Catherine Thome

Robert J. Tierney | Sandra Wilde

HEINEMANN

Portsmouth, NH

Heinemann
A division of Reed Elsevier Inc.
361 Hanover Street
Portsmouth, NH 03801–3912
www.heinemann.com

Offices and agents throughout the world

A previous edition of this book, titled *Examining DIBELS: What It Is and What It Does*, was published by the Vermont Society for the Study of Education.

Published in cooperation with the Vermont Society for the Study of Education (VSSE) and the Center for the Expansion of Language and Thinking (CELT).

Library of Congress Cataloging-in-Publication Data
The truth about DIBELS : what it is, what it does / edited by Kenneth S. Goodman ; foreword by P. David Pearson.
p. cm.
Includes bibliographical references.
ISBN-13: 978-0-325-01050-2
ISBN-10: 0-325-01050-1 (pbk. : alk. paper)
1. Reading (Primary)—Ability testing—United States. 2. Educational tests and measurements—United States—Evaluation. I. Goodman, Kenneth S.

LB1525.75.T78 2006
372.48—dc22 2006014782

Acquisitions editor: Lois Bridges
Editor: Gloria Pipkin
Production: Lynne Costa
Cover design: Night & Day Design
Typesetter: House of Equations, Inc.
Manufacturing: Louise Richardson

Printed in the United States of America on acid-free paper
10 09 08 07 06 VP 1 2 3 4 5

Contents

Foreword

P. David Pearson

*W*hen it comes to controversial issues in the teaching of reading, I have built a reputation for taking positions characterized as situated in "the radical middle" (Pearson, 2001)—not too conservative, not too liberal, just right (or at least I like to think so!). Not so on DIBELS. I have decided to join that group of scholars and teachers and parents who are convinced that DIBELS is the worst thing to happen to the teaching of reading since the development of flash cards.

I take this extreme position for a single reason—DIBELS shapes instruction in ways that are bad for students (they end up engaging in curricular activities that do not promote their progress as readers) and bad for teachers (it requires them to judge student progress and shape instruction based on criteria that are not consistent with our best knowledge about the nature of reading development).

The appeal of DIBELS. So if a group of eminent scholars such as those represented in this volume thinks DIBELS is so terrible, so malevolent, and so harmful, then why is it so popular? Why is it used in so many states, districts, and schools? Several reasons come readily to mind—some technical, some curricular, and some political in origin.

First, DIBELS has tremendous scientific cachet. If you go onto the DIBELS website, you find yourself awash in a sea of numbers—reliability indices, validity indicators, the number of students currently using DIBELS (almost 2,000,000 at the latest count). I have included a few of the more important ones in Table 1, which I extracted from the publicly accessible data on the website (http://dibels.uoregon.edu).

Table 1: *Psychometric characteristics of DIBELS*

	Alternate Form Reliability	**Criterion-related Validity**	
Letter Name Fluency	.88	Concurrent Validity WJ: .70	Predictive Validity .65–.71
Initial Sound Fluency	.72	DPSF: .48 WJ Readiness: .36	CBM: .45 WJ Reading: .36
Phoneme Segmentation Fluency	.79	WJ Readiness: .54	DNWF: .62 WJ Reading: .68
Nonsense Word Fluency	.83	WJ Readiness: .59	CBM: .60–.82 WJ Reading: .66
Oral Reading Fluency	.89–.94*	.52–.91*	
Retell Fluency	.59: ORF*		

KEY: WJ — Woodcock-Johnson
CBM — Curriculum-Based Measures
DPSF — DIBELS Phonemic Segmentation Fluency
DNWF — DIBELS Nonsense Word Fluency

*The estimates for the reliability and validity of the ORF approach are based upon older studies documenting the general approach of Curriculum-Based Measures (Good & Jefferson, 1998; Tindal, Marston, & Deno, 1983) as cited on the DIBELS website (Good & Kaminski, 2002), not on the specific passages included in DIBELS.

From a psychometric perspective, the reliability data are impressive, especially for individually administered tests requiring human judgments about response correctness. One can trust the scores to be stable, at least in the short run (see Paris, 2005). And the validity indicators tell us how much these tests are similar to other tests of reading and verbal ability. I have divided the construct of criterion-related validity into two categories: (a) *concurrent* (how well does a given subtest correlate with scores on a test given at the same time as the subtest?) and (b) *predictive*

(how well does a subtest predict scores on a test given at some point in the future?). What can be said of these correlations is that they are roughly of the same magnitude that we find among a wide range of measures of reading and verbal measures (see Paris, 2005, for an account of these patterns of covariation). It is interesting to note that the psychometric data for the Oral Reading Fluency and Retelling measures are based on the assumption that evidence about the general pool of Curriculum-Based Measures of which the DIBELS passages are a part will suffice as a measure of their psychometric rigor. On this matter, it is important to note that in a recent independent study of the predictive validity of DIBELS Oral Reading Fluency (given in Grade 3 and used to predict end of the year scores on the Terra Nova standardized reading test), Pressley et al. (2005) concluded that ". . . DIBELS mis-predicts reading performance on other assessments much of the time, and at best is a measure of who reads quickly without regard to whether the reader comprehends what is read."

Another reason for the appeal of DIBELS is its transparent match with the sort of curriculum championed by the Reading First plank of No Child Left Behind (2002). This association links DIBELS to another "scientific" indicator, the National Reading Panel (NRP) report (2000) by virtue of the fact that the NRP report serves as the research architecture for Reading First. There can be no doubt that the NRP's "big five" (phonemic awareness, phonics, fluency, vocabulary, and comprehension) shape instruction in schools with Reading First grants, and by extension, in *all* schools (because states, quite understandably, want to ensure alignment to NCLB/RF, and the scientific aura that comes with it, for all their schools, not just the ones eligible for extra funding).[1] It is also interesting to note that

[1]While it goes beyond the scope of this essay, it is important to note that there is nothing magical about the focus on this particular set of "big five" components of reading. The impact of writing or oral language development on reading, grouping, text difficulty, talk about text, and opportunity to read are just as important a set of curricular components as are the current big five, but they did not surface in the NRP report (2000), either because the body of experimental research undergirding them was insufficient to permit either meta-analysis of the sort championed by the panel or because these issues lay outside the boundaries of the individual and

DIBELS is available in conjunction with certain commercial programs; for example, Scott-Foresman (in the spirit of full disclosure, the company with which I serve as a basal author) markets DIBELS alongside of its reading program for elementary kids. This sort of juxtaposition only increases its transparent connection to curriculum.

The third reason for DIBELS' appeal is simplicity and ease of use. Imagine getting important information for monitoring student progress with one-minute samples of performance. Not only how many words read correctly in a minute, but rates for every possible behavior: letters named per minute, phonemes identified per minute, words named per minute, words recalled from a passage in a minute. Contrast the ease and simplicity of DIBELS with the exhortations of nuance and complexity one gets with the detailed analysis of oral reading one gets from running records or miscue analysis or from schemes for understanding students' retelling. It is easy to see how and why a busy teacher, overwhelmed with the responsibilities of planning instruction for students with many needs and meeting the requirements of bureaucratic reporting systems, might opt for the efficiency of DIBELS.

The fourth plank of DIBELS' appeal is based on its political positioning in the enactment of NCLB and Reading First policy. Evidence providing direct links between the advice or mandate of federal officials and the tools chosen by various states as the official scientifically based assessment portfolio for their Reading First implementations is always difficult to document (but see *Education Week*, September 7, 2005; Manning, Kamii, & Kato, this volume; Manzo, *Education Week*, September 28, 2005, for reports of such influence). Nonetheless, the ubiquity of DIBELS

collective interests of the panel members (Samuels, 2006). Researchers have no responsibility, in principle, to ensure that all aspects of a full reading curriculum are addressed in such a review. For better or worse (probably better) teachers and schools do. They have to decide what texts kids should read even though the research on optimal text types or combinations thereof do not permit a definitive conclusion about what kids should read. The danger, of course, with a document like the NRP is that the education profession will be seduced into believing, even though the NRP never intended so, that the big five is *all* that schools and teachers need to worry about. Even a cursory examination of state standards and state programs to guide the use of Reading First funds would suggest that the narrow view of curriculum promoted by the big five is real, not imaginary.

as the "preferred" progress monitoring tool across various states gives one pause to wonder whether informal coaching, in some cases out and out insistence on the part of federal officials or reviewers (Manning, Kamii, & Kato, this volume) to move in the direction of DIBELS, did not play a more important role than natural market forces in creating the obvious competitive advantage it holds (*Education Week*, September 7, 2005; Manzo, *Education Week*, September 28, 2005). Of course, it does not hurt that DIBELS was officially blessed as a "scientifically valid" instrument for purposes of progress monitoring by the Reading First Assessment Academy (http://idea.uoregon.edu/assessment/index.html), an advisory group on which DIBELS author Roland Good served. It seems to me that federal officials, at a minimum, are ethically and professionally obligated to explain the widespread use of DIBELS (especially in a marketplace flooded with potential competitors) and the apparent conflict of interest that seems to occur in this situation. I would even appreciate it if someone had the courage to 'fess up what seems, on the face of it, to be such an obvious conclusion—that privilege and favoritism, not merit, are behind this unlikely dominance of DIBELS across such a wide range of states.

The Problem with DIBELS

So what is the problem with DIBELS? Why am I, and why are so many other scholars, so concerned about its widespread and ever-increasing use? The answer to this question is, of course, what this book is all about. In the anchor paper in this book, Ken Goodman argues convincingly that DIBELS is much more than a test—that it has, de facto, become an implicit (perhaps even an explicit) blueprint for a curriculum—driving publishers, district officials, principals, and teachers into a narrow curricular mode in which only the big five (and mostly the even bigger three of phonics, phonemic awareness, and fluency) are taught at the expense of other curricular foci. The cost, in terms of human frustration and curricular opportunity, Goodman argues, is serious.

The other papers in this volume—by Tierney and Thome on its broad policy influence; by Seay on its failure to deliver its promise of increased achievement to Alabama; and Manning, Kamii, and Kato on its concurrent validity—all make related points about DIBELS. Interestingly, all either assume or draw similar conclusions about DIBELS' capacity to shape instruction in counterproductive ways by directing schools and teachers to a limited set of features of the reading curriculum. Most directly on point is the paper by Tierney and Thome, in which they point out the costs that schools, teachers, and students must bear when means are confused with ends. Students are held accountable to the indicators rather than the outcomes of progress. Teachers are forced to shape instruction in ways that violate well-documented theories of development, to privilege some aspects of literate performance over others, and to value students' performance over their vitality and identity. More importantly, Tierney and Thome point out, in a stroke of irony, that teachers are forced to assume a professional disposition at odds with the document serving as the architecture for Reading First—the National Reading Panel report (Tierney & Thome, this volume, page 50).

Of course, the authors of DIBELS, and the Reading First Assessment Academy that blessed DIBELS, would argue that these stories of curricular influence represent misuses rather than valid uses of DIBELS. They would say that DIBELS is authorized for only one of four primary purposes of assessment, progress monitoring, and, as such, it should be used only as a "thermometer" to gauge student learning, not as a "vision" to guide specific instructional foci.[2] But one of the first statements one encounters on the DIBELS website (http://dibels.uoregon.edu/dibelsinfo.php) is this description of what they are and what they are good for:

> The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) are a set of standardized, individually administered measures of

[2]The four purposes of assessment required in Reading First are (a) screening—determining who might need extra help, (b) progress monitoring—benchmark assessments administered at regular intervals to determine who is, and is not, on track, (c) diagnosis—determining specific needs of specific students to guide specific interventions, and (d) outcome assessment—judging the effectiveness of a program, intervention, curriculum, etc.

> early literacy development. They are designed to be short (one minute) fluency measures used to regularly monitor the development of pre-reading and early reading skills.
>
> The measures were developed upon the essential early literacy domains discussed in both the National Reading Panel (2000) and National Research Council (1998) reports to assess student development of phonological awareness, alphabetic understanding, and automaticity and fluency with the code. Each measure has been thoroughly researched and demonstrated to be reliable and valid indicators of early literacy development and predictive of later reading proficiency to aid in the early identification of students who are not progressing as expected. When used as recommended, the results can be used to evaluate individual student development as well as provide grade-level feedback toward validated instructional objectives.

The major unknown in this formulation is, of course, what is done with the information provided to fulfill the goal of the last sentence: ". . . to evaluate individual student development as well as provide grade-level feedback toward validated instructional objectives." If the assumption is that teachers should teach and students should learn the skills measured by DIBELS, then these subtests do, in fact, become a curricular blueprint. One could imagine a parallel world in which they really were used as a thermometer—where a full and balanced curriculum was provided, and speed and fluency were regarded as the natural outcomes, not the objectives, of such a curriculum. My own view: in most (not all, but most) places in which it is used, DIBELS guides instruction right into the big five (and in lots of places the even bigger three), and all else (e.g., writing, oral language, disciplinary knowledge, discussion) has to compete for a very small piece of the curricular pie.

Of parallel importance, Goodman argues, is that DIBELS is based upon a flawed view of the nature of the reading process and, because of this fundamental flaw, provides all who use it with a misrepresentation of reading development. It digs too deeply into the infrastructure of reading skill and process and comes up with a lot of bits and pieces but not the orchestrated whole of reading

as a skilled human process. Manning, Kamii, and Kato, in their chapter, provide correlational evidence to support this point, noting that the Phoneme Sound Segmentation Fluency test demonstrated only the most modest of correlations with a concurrently administered invented spelling task (a task in which phonemic awareness is absolutely essential) or with the Slosson Oral Reading Test (r = .07).

I want to add one more item to the "what's wrong" list, one alluded to in the current book—but one that concerns me so gravely that I want to use this opportunity to unpack it. This criticism focuses on what I take to be a sort of psychometric alchemy, and it turns on the metrics that DIBELS uses to scale student performance—rates (the number of X per minute) rather than accuracy (percentage of this domain the student exhibits control over). Scott Paris, in a hard-hitting critique of the ways in which lots of these tests of specific skills are used (2005), makes an important distinction between what he calls constrained and unconstrained skills. This distinction parallels one that I borrowed from the late James Squire, between mastery skills and growth constructs. Paris's constrained skills (my mastery skills) are phenomena—such as letter names, letter sounds, phonemic awareness, mechanics in writing—that we teach with the expectation that once kids demonstrate mastery, we can get on with something else. Paris's unconstrained skills (my growth constructs) are phenomena—such as composition, comprehension, word meaning, or critical thinking—that, by their very nature, cannot be mastered; they always exhibit capacity for even greater growth. We teach them not with the expectation that they get learned to mastery so we can go on to something more important, but with the expectation that they are the real stuff of literacy—the important things we go on to!

If one looks at DIBELS, it is clear that all of the tests, except the Oral Reading Fluency/Retell component, measure constrained (mastery) skills—things that are learned along the way to real reading. If the creators of DIBELS had chosen to report these phenomena on an accuracy scale rather than a rate scale, what they would have found is that for many of the skills, most

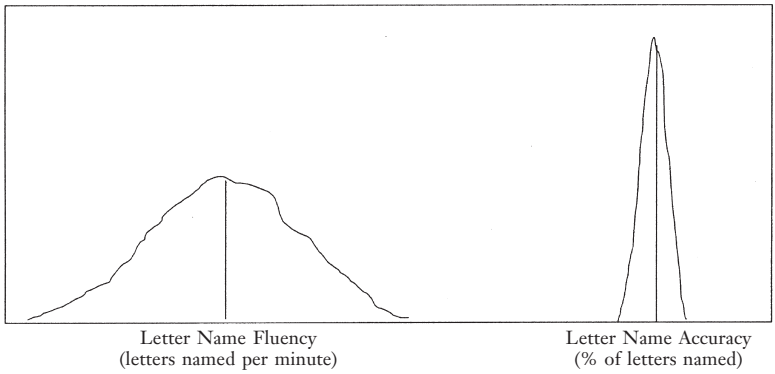

Figure 1: *Hypothetical comparative distribution of fluency and accuracy scores on letter naming*

students would reach a performance ceiling in either first or second grade. Such a distribution of scores is depicted in the right-hand side of Figure 1. But that same set of students who vary little in terms of the alphabet knowledge will vary considerably in their letter naming fluency (letters named per minute), as depicted in the left-hand side of Figure 1.

There is no magic, no alchemy, here. It is a direct application of a general rule about human performance: human beings alike in their capacity to perform a given task with accuracy and integrity will vary dramatically in the speed or fluency with which they perform it. This is true not only in reading but also in a range of cognitive and psychomotor tasks. It is part of our "human nature," if you will. It is also the case, again in a wide range of human performances, that speed or fluency correlates with general cognitive and psychomotor ability. The creators of DIBELS benefit from these two general laws. It permits them to show that performance on a highly constrained skill on which students perform at near mastery levels predicts scores on a more general unconstrained phenomenon (such as a standardized reading test score) if the accuracy index on that constrained skill is transformed into a fluency index. Why? For a lot of reasons, but directly to the point here is a long-standing rule of thumb in testing: other things being equal, the greater the variance on a test, the greater its reliability and, hence, the greater its capacity to demonstrate cor-

relations with other measures of cognitive ability. This is what I was referring to earlier when I used the metaphor of psychometric alchemy. DIBELS, and any other test battery that uses rate rather than accuracy as an index of performance for a constrained skill, can demonstrate the concurrent or predictive validity of that measure to a greater degree than can a test battery that relies on accuracy.

The important question for a teacher or a school faculty is the "so what" question: What will you do differently instructionally if you know that Tommy names his letters or his phonemes at a slower rate than does Susana? Will we have Tommy practice naming letters faster? The same could be asked of the ubiquitous reading fluency measures we see all over America. What will we do with students who read accurately but slowly—have them engage in timed trials five days a week? Now we are back to the means-ends confusion discussed earlier—the problem that arises when an indicator of progress is elevated to the status of a curricular goal. I surely want kids to make progress on fluency—to read faster and with greater expression—as they mature (by the way, I also want them to know when to slow down—either to savor the language of an author or to puzzle over an enigmatic expression) as Flurkey demonstrates in his chapter. But I want them to make that progress on fluency because we provided them with a rich curriculum that ensured balanced development of a range of skills and a broad exposure to important ideas—not because we had them practice timed trials five times per week.

Changing My Mind

What would it take to convince me to change my mind about DIBELS? I like to think of myself, ultimately, as the quintessential empiricist—as a person who would be convinced by evidence that my a priori conceptual stand was misguided or just plain wrong. I also like to think of myself as an individual who has changed his mind often—whenever compelling evidence and logic forced me to reconsider my position and/or world view. To those unhappy, even outraged, by my critique, I do have an experiment

in mind for DIBELS proponents and the folks in Reading First and No Child Left Behind who have, either directly or indirectly, promoted its use. It would require the DIBELS folks to move beyond criterion-related validity to validity related to the consequences of test use (Messick, 1989) to demonstrate that when DIBELS is used in the ways that its proponents suggest, good things happen to and for students, teachers, and parents. Messick coined the term *consequential validity* to capture the notion that the validity of tests is less about their internal character than about the decisions that are made when they are used to shape action in the real world. So here is my challenge to the DIBELS folks: Show me that when DIBELS is used to monitor and shape instruction (as it clearly is in the current milieu) it actually promotes growth on more significant and more global indicators of reading (and writing) development than are measured by DIBELS itself. If, as a result of using DIBELS to guide instruction, kids read more, read more *enthusiastically* and with *greater comprehension*, wrote with greater *facility*, and *felt better* about themselves as readers, then I would back off this critique and say, "You're right. Using DIBELS does help develop more avid, active, and efficacious readers and writers than other assessment tools."

In my experiment, there would be two large groups (with lots of relevant sub-populations): one using DIBELS and one using more global indicators to monitor progress and shape instruction. The two progress-monitoring assessments would be given at key points along the way, probably three times a year over a three- or four-year period (I'd propose grades K–3). At prespecified intervals, we'd also administer more global, more growth-oriented assessments of reading and writing, at a minimum at the beginning of the study and at the end of each academic year. The key question would be whether teaching in a way to directly influence growth on DIBELS promoted incidental growth on the global measures to a greater degree than did teaching for growth with some alternative monitoring system. In other words, it would answer the question of whether DIBELS bore more positive consequences for kids than did some alternative. And if it did, I'd back off the critique I have

offered—because the data would have demonstrated that the path to literacy is paved with letters, phonemes, and fluency, not, as I currently believe, with richer knowledge, more refined word meanings, and an extensive tool kit of useful strategies to make sense of text when the going gets tough. In this regard, I would note that Seay (this volume) does provide a "partial" answer to this question, albeit at a very broad level of analysis. She examined the changes in scores of Alabama students on NAEP reading and on the state's standardized reading test (SAT-10) as a function of either 2 or 3 years of participation in a Reading First program in which DIBELS was used to monitor and guide instruction. She notes little improvement on either of these external indicators as a function of Alabama's participation in Reading First (but it must be admitted that she did not disaggregate state assessment performance as a function of whether schools did or did not participate in Reading First). Even so, the anticipated gains in overall achievement have not been realized.

Coping in the Meantime

Finally, there is the question of what we do while we are waiting for this "millennium" study to be done. How do we cope with the reality of DIBELS? Here is where this book will be useful, for it outlines strong critiques that concerned educators and citizens can take to school boards and legislatures and suggests alternative models of responsible assessment—what could, and perhaps should, we be using to monitor student progress.

If I were working in a district or school where DIBELS was mandated, I'd insist that we develop and implement a set of parallel assessments that measure reading and writing in their more global, not their more atomistic, aspect—maybe something like running records with comprehension and response to literature, regular writing samples, and some index of spelling progress. And if kids were making progress on DIBELS but not the more global measures, I'd want to argue for a different sort of intervention than is typically promoted by DIBELS. And if kids were making progress on the more global measures but not DIBELS,

I'd want to know the sorts of compensatory mechanisms they were using in the absence of well-developed alphabetics.

The final thing I would do is to promote, at every opportunity, greater sensitivity to what I consider the two most important principles of good assessment policy. Principle #1 addresses the issue of how assessment relates to curriculum and suggests that we beware of putting the cart before the horse. The point is that assessments should reflect, not lead, curriculum and instruction. We need instructionally sensitive assessments, not assessment-sensitive curriculum.

> Principle #1: Never send a test out to do a curriculum's job!

The second principle relates to the question of consequences very directly. It admits that, other things being equal, people will teach to tests—even if in their heart of hearts they know they should not. Further, it suggests that the higher the stakes (consequences), the greater the likelihood that people will teach to a test. Hence, when stakes are high, so must be the level of challenge and the transparent authenticity of the test.

> Principle #2: The higher the stakes, the greater must be the challenge and the authenticity of the assessment.

The worst situation imaginable is high stakes and low challenge—for that combination will drive instruction to the lowest common denominator and guarantee that our lowest achieving students will never get to the "good stuff" in our curriculum because they will spend all of their time working on the "basics."

Now to the Book

As I said at the outset, I agreed to write this foreword because I think the crisis of curriculum promoted by excessive reliance on componential measures of reading such as DIBELS is serious—palpable, you can feel it and almost touch it—in our schools when you view the consequences and hear the stories of parents and teachers and students whose lives are directly influenced by these assessments—and by DIBELS in particular. The book is impor-

tant to those who wish to resist this sort of curricular mandate because it offers strong arguments and evidence to support that effort. It is my hope that the book will also spur reading and measurement researchers to pursue rigorous research to address some of the unanswered questions that remain before us regarding the conceptual, psychometric, and pragmatic aspects of assessment policy. In fact, I would hope that the federal Department of Education would seize the opportunity to ensure that the assessments it promotes are held to the gold standard of assessment scholarship and fund studies to evaluate the validity and impact of various sorts of assessment tools, including DIBELS.

Happy reading on the road to action in the policy arena—at every level: local, state, and national.

References

Good, R. H., & Jefferson, G. 1998. *Contemporary Perspectives on Curriculum-based Measurement Validity*. New York: The Guilford Press.

Good, R. H., Kaminski, R. A. 2002. *DIBELS Oral Reading Fluency Passages for First Through Third Grades* (Technical Report No. 10). Eugene, OR: University of Oregon.

Manzo, K. 2005. "National Clout of DIBELS Test Draws Scrutiny." *Education Week* 25 (5): 1, 12.

Manzo, K. 2005. "States Pressed to Refashion Reading First Grant Designs." *Education Week* 25 (2): 1, 24–25.

Messick, S. 1989. "Validity." In *Educational Measurement*, 3rd ed., edited by R. L. Linn. New York: Macmillan.

National Institute of Child Health and Human Development. Report of the National Reading Panel. 2000. *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.

Paris, S. G. 2005. "Reinterpreting the development of reading skills." *Reading Research Quarterly* 40 (2): 184–202.

Pearson, P. D. 2001. "Life in the radical middle: A personal apology for a balanced view of reading." In *Reading Researchers*

in Search of Common Ground, edited by R. Flippo. Newark, DE: International Reading Association.

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*. East Lansing, MI: Literacy Achievement Research Center, technical report.

Tindall, G., Marston, D., & Deno, S. L. 1983. *The Reliability of Direct and Repeated Measurement* (Research Rep. No. 109). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Prologue
DIBELS: One Family's Journey

Lisa Laser

I can think of no better prologue to this critique of DIBELS than this diary of one family's experience with DIBELS.
KSG

Note 1: Mid-September, 2005

I've been crying for the better part of two days because my family has encountered the horror of DIBELS. My husband and I just moved from Portland, Oregon—where our son, Ellis, had a wonderful kindergarten experience—to a rural town in the northeastern section of the state.

Ellis is starting first grade at Joseph Elementary. We were surprised but not alarmed to learn from the first grade teacher that all incoming first graders (according to her) knew how to read. Ellis does not know how to read, nor do we support a teaching philosophy that would make reading mandatory for entering or even finishing first grade.

After teaching Ellis for thirteen days, the first-grade teacher at Joseph Elementary asked us for an immediate conference to discuss his progress. She gave us two choices: Ellis could repeat kindergarten or could be held back at the end of first grade. These options were presented to us as the best solution for Ellis to keep him from being "stigmatized because he was below the benchmark in two specific DIBELS scores, phoneme recognition and nonsense words. (Ellis was average to above-average for his class in the other DIBELS areas. She did not address the issue of other subjects such as math and science where we are confident he'd

be ahead of his class since his Portland school had a math and science focused curriculum and we look over his daily school work when he arrives home.

The main issue for us is Joseph Elementary's overreliance on standardized testing and teaching methods solely designed to support test success—the child's well-being be damned. Another issue is the lack of respect and support for students who move across state and change schools. Whatever happened to extra support, such as tutors or make-up lessons conducted at home? The school staff did not offer these solutions; additionally, we received teaching materials only after we requested them.

Unfortunately, the only other public school nearby also uses DIBELS. We are currently debating the option of homeschooling and supplementing our son's education to make up for what is left out of the public school curriculum in order to focus so heavily on standardized tests. This trade-off has led to a disappearance of creativity, diversity, philosophical inquiry, and fun in our public schools; it seems so monumental and ingrained in the current system that we are sickened at the thought of keeping Ellis in such an environment. We will not be holding back our bright, confident, tender-hearted, funny, and delightful son.

Note 2:

One of my favorite music groups is the Violent Femmes (I experienced high school in the early 80s). The song that used to make us all scream at the top of our lungs was "Kiss Off"; it came back to me today in all its brilliance for one particular line:

> I HOPE YOU KNOW THIS WILL GO DOWN ON YOUR PERMANENT RECORD.

At Joseph Elementary (grades K–4) the students are tested approximately once a month by the Title I coordinator. These scores are then transmitted to the University of Oregon.

Ellis' first test had the following scores; the numbers in parentheses represent the range of the fifteen other students in his class:

Letter Naming Fluency: 28 (11–45)

Phoneme Segmentation Fluency: 4 (4–68)

Word Use Fluency: 41 (14–63)

Nonsense Word Fluency: 6 (6–44)

No one at Joseph Elementary ever discussed DIBELS and the monthly testing when we enrolled Ellis. No one has ever discussed the transmission of his test scores either. Welcome to *1984*.

About the nonsense word portion of DIBELS: Ellis is a very thoughtful and cautious boy, so careful and deliberative is exactly how he would approach a test. He is not one to be rushed, whether it's getting his shoes on or painting a picture.

I would like to report that the first-grade teacher is young and inexperienced, but, in fact, she is neither—my guess would be she's been teaching at least twenty-five years. She also seems to truly believe in DIBELS, which is more horrifying than those who are trapped by it and NCLB.

I spent time in Ellis' classroom yesterday; Bob and I will be going next week as well. What I saw makes me unable to give my son to such a rigid and unimaginative place. There was nothing redeeming about his day. Not even art was fun.

I've met with people homeschooling their children nearby. At this point, that is our plan.

Note 3: Oct. 26, 2005

September 14th, the day we met with Ellis' teacher to discuss his progress, was the beginning of a whirlwind education on DIBELS. That evening, I spent about five hours reading everything I could find. I plan to continue to fight whether or not my son is in school. But Ellis' teacher wholeheartedly believes in DIBELS and teaches to the test for forty-five minutes, twice a day.

During Ellis' first two weeks at Joseph Elementary, he was pulled out for extra instruction every day. He was an absolute mess—tired, cranky, short-tempered—and he was becoming depressed.

On Monday, October 7, Ellis came to the decision on his own to not go to school any more. That day, he seemed especially tired and depressed. Bob and I had already decided that sending him back to school was detrimental to all of us. We had wrestled with how to tell Ellis he wasn't going back to school; for days we just said we wanted him to take time off. And then on Wednesday, he was having one meltdown after another, so I took him aside at a quiet area in the park. Ever since he'd been going to Joseph Elementary he seemed tired, grumpy, and not quite himself, and I asked him what he thought we could do to help him. He said, "Stop sending me to school." We can sure support that wish!

When we then met with the principal to discuss why we were withdrawing Ellis, she gave us the Benchmark Assessment used to test him. We found out that he was tested first on August 26th and again September 28th. It was infuriating to know that even with the improvement shown in that short period of time, Ellis' teacher wanted him to repeat kindergarten. It was clear to us that DIBELS is double punishment: DIBELS, in and of itself, and the expectation that kindergartners must enter first grade with a specific and narrow range of reading skills in spite of their other skills and knowledge.

Getting an inside look at the test was additional proof of how lousy DIBELS is. For example, the font used in the letter naming fluency section was one where the lower case *g* and *a* look alike, which is neither the way children are taught to write them nor the font used in children's books. Also, the letter *l* looks just like the number one, further confusing students. Ellis' test twice indicated that he called the *l* a *1*—or was it the *1* an *l*?

We have been homeschooling now for four weeks and Ellis' change back to his regular lovely self is astounding, although not surprising after witnessing his school day. We are not willing to sacrifice our child to that hellhole. I will, however, continue to do what I can to change what is happening to our public schools because of DIBELS.

But Isn't DIBELS Scientifically Based?

Sandra Wilde

*I*t suddenly seems as though DIBELS testing is everywhere; in fact, it seems to have replaced other assessments of reading in many school districts, particularly those receiving Reading First grants. Those who endorse the use of DIBELS say that one of the program's strong points is that the tests are an accurate predictor of potential reading failure—indeed, teachers have reported to me that this is what they are told in DIBELS workshops. The implication is that even if the DIBELS procedures seem unconnected to what real reading looks like, they're valuable in providing a quick screening tool for identifying which students need extra help.

I recently decided to put this claim to the test. I examined the findings from two studies linked to on DIBELS's own website at the University of Oregon. Since these studies have been made available by the program itself, I assumed they would offer a best-case scenario of how strong a predictor DIBELS is. Both studies deal with the Oral Reading Fluency subtest (see the Appendix), the subtest most often used and indeed the only one offered in grades 2–6.

First, I looked at a study by Buck and Torgesen (2003), which aimed to find out "whether performance on brief, one-minute measures of oral reading fluency are predictive of achievement in reading as measured by the reading portion of the [Florida state assessment]." The students were categorized by the DIBELS Oral Reading Fluency subtest (ORF), which measures the number of words read correctly in one minute, into three groups: high, some, or low risk. Each of these groups was then divided into two sub-

Table 1: *Buck and Torgesen Study Data*

	Words correct per minute			
	High Risk: <80	Some Risk: 80–109	Low Risk: >110	Totals
Adequate (Meeting State Benchmark)	42	188	511	741 (67%)
Inadequate (Not Meeting State Benchmark)	178	130	53	361 (33%)
Number and Percent of Total	220 (20%)	318 (29%)	564 (51%)	1102

Sensitivity = .77
Specificity = .92
(Calculated only for the high-risk and low-risk groups)

groups, those who did and didn't meet the state benchmark in reading. The DIBELS tests were conducted a month after the state tests, so what's being measured is not so much prediction of a future test score but correlation with a current one, which statistically amounts to pretty much the same thing. The results appear in Table 1.

At first glance, it seems that the DIBELS test was a pretty good predictor. After all, students who scored as low-risk on it were likely to meet the state standard, while high-risk students were not. However, let's think about what the numbers might mean for actual classroom practice.

If this DIBELS test were used to decide which students should receive extra instruction, schools would have to determine whether to help just the high-risk students or the some-risk ones as well. There were 361 students who didn't meet the state benchmark; only half of them (178) would have gotten extra help if it only went to those in the high-risk category. If help were given to the high-risk and some-risk categories, it would catch most of those likely to not meet benchmark (308), but extra help would also be given to 230 students who didn't need it. Indeed, the high- and some-risk groups combined make up nearly 50 percent of the study's population.

Table 2: *Wilson Study Data*

	High Risk	Some Risk	Low Risk
Meeting State Standard	7.0%	51.4%	81.9%
Not Meeting State Standard	93.0%	48.6%	18.1%

Let's look at another study, this one from Arizona (Wilson 2005). The tests were also given to third graders, in this case 241 of them. The study didn't indicate how many students fell into each subgroup, so their results, in Table 2, appear as percentages only. (48 percent of the students passed the state test.)

Again, the prediction is strong only if you leave out the "some risk" group; half the students in this group passed the state assessment and half didn't, so the DIBELS test had no predictive value at all for them. One would again be in a situation of providing extra help to students who didn't need it or missing those who did.

What conclusions do the authors of these studies draw? According to Buck and Torgesen, "This initial study demonstrates that, for a large heterogeneous group of third graders, performance on brief oral reading fluency measures can quite accurately predict whether or not a given student will attain a [passing score on the Florida] reading test" (2003, 9). Well, no. Only those who score in the high-risk or low-risk group, which of course include the strongest and weakest of readers, those whose teachers already know if they're likely to pass the state test. Interestingly, Buck and Torgesen chose to leave out the some-risk group in calculating sensitivity and specificity scores—roughly, how likely the results are to be accurate: the true positive and true negative, respectively. The test isn't really predictive at all for the middle, some-risk group.

Wilson's conclusions are also problematic. It may be reasonable to say, as he does, that "ORF can identify those students who are likely to meet the proficiency standard on [the state test] with good accuracy" (i.e., 82 percent), but it's not true that it can "identify those who are quite unlikely to reach proficiency" (2005, 4).

Again, no. Not unless you also include quite a few students who don't need help.

So much for the vaunted claims that a one-minute read-aloud is all that you need to identify who needs extra help with reading. I've only looked at two studies here, but is it likely that other studies would show dramatically different results? Although many schools use DIBELS as an initial prescreen to be followed by other assessment, Oregon teacher Donna Shrier (personal communication 2006) informed me that in her district, the DIBELS oral reading score determines whether students are in Title I programs and what kind of instruction they get. Teachers are required to provide any student who performs at the high-risk level with intensive instruction that is heavily phonics-based, "because that's the only way these kids can learn." Another teacher in the district told me that a student of hers who performed at the benchmark level on DIBELS (i.e, in the low-risk range) was removed from Title I instruction even though she hadn't passed the state test and, in the teacher's opinion, needed extra help. When I asked her what the district office would say if asked the rationale for this policy, she replied, "They'd say it's scientifically based."

Well, maybe this scientific emperor has no clothes, or skimpy ones at best. The DIBELS website supports the Oral Reading Fluency subtest by saying that eight studies in the 1980s found criterion-related validity of the test (usually determined by comparing a measure with other tests) to range from .52 to .91 (dibels.uoregon.edu/measures/orf2.php), hardly a stunning performance when it comes to making decisions about individual students, especially upon taking a closer look at the numbers as I've done for the two studies described.

And it's the effect on individual students that really matters here. It would be one thing to use DIBELS to predict, for instance, what proportion of students in a school are likely to fail the state reading test, but it's clearly not that good at identifying *which* students they'll be. Also, if DIBELS were used as a quick initial screen to identify which students need further assessment, that would be reasonable (although it would still miss some cases),

but it often drives out other forms of assessment that would be far more sensitive to a variety of reading problems. The hard sell of DIBELS promotional efforts, with a strong veneer of being "scientifically based," has had a powerful effect on educational decision makers.

I haven't even tried to touch on the larger issues, such as whether state reading tests are a good measure of reading competence, let alone what kind of instruction best benefits struggling readers. But DIBELS fails on the most basic grounds of validity; that is, whether it measures what it claims to be measuring. As Kenneth Goodman stated earlier in this volume, scores on reading tests tend to correlate highly with each other no matter what. But the DIBELS Oral Reading Fluency subtest claims to do something more: to strongly predict whether individual children are likely to fail to learn to read. It just doesn't.

References

Buck, J., and J. Torgesen. 2003. *The Relationship Between Performance on a Measure of Oral Reading Fluency and Performance on the Florida Comprehensive Assessment Test*. FCRR Technical Report #1. Tallahassee, FL: Florida Center for Reading Research. dibels.uoregon.edu/techreports/index.php.

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)*. Tempe, AZ: Tempe School District, Assessment and Evaluation Department. dibels.uoregon.edu/techreports/index.php.