1-1-2007

Impact of test-wiseness training on kindergarteners' performance on early reading measures

Keath A. Murray
Eastern Illinois University

Impact of Test-Wiseness Training on Kindergarteners' Performance on Early Reading Measures

BY

Keath A. Murray

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

Specialist in School Psychology

IN THE GRADUATE SCHOOL, EASTERN ILLINOIS UNIVERSITY
CHARLESTON, ILLINOIS

2007
YEAR

I HEREBY RECOMMEND THIS THESIS BE ACCEPTED AS FULFILLING
THIS PART OF THE GRADUATE DEGREE CITED ABOVE

6/28/07
Date

Christene McCormick
Thesis Director

6/28/07
Date

Department/School Head

Running head: TEST-WISENESS AND EARLY READING MEASURES

Impact of Test-Wiseness Training on Kindergarteners' Performance on Early Reading Measures

Keath A. Murray

Eastern Illinois University

Abstract

The purpose of this study was to examine the effect of test-wiseness (TW) training on kindergartners' performance on an early reading measure. Scores obtained from kindergarten students who receive TW training for two of the kindergarten probes from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) were compared to scores of kindergarten students did not receive TW training. Approximately 52 kindergarten students were randomly assigned to 2 conditions. In one condition, 26 students took an initial probe, received brief TW training 2 weeks after the initial probe followed by a second, alternate-form "practice probe", and then a third alternate-form probe. In the other condition, 26 students took an initial probe, followed by an alternate-form "practice probe" 2 weeks after the initial probe with no prior TW training, and then a third alternate-form probe. A t-test was conducted to compare the mean LNF and NWF scores of the TW group to the control group, and effect sizes were calculated using Cohen's *d*. The study's hypothesis was that those students exposed to TW training prior to the actual testing will achieve significantly higher test scores than those who were administered the DIBELS in its traditional fashion

Acknowledgements

This study was conducted under the supervision and advisement of Dr. Jason Nelson. Dr. Nelson provided guidance in research and implementation of this study as well as in the analysis of the research data. The researcher would also like to acknowledge the members of the thesis committee for their guidance throughout the development and implementation of this study. Finally, this research would not have been possible without the help from the participating elementary school and those involved in the implementation of the study.

Impact of Test-Wiseness Training on Kindergarteners' Performance on Early Reading Measures

A push toward a problem-solving approach has been growing within the field of school psychology in response to criticism of the IQ-achievement discrepancy model of learning disabilities (LD) identification. With the passage of Individuals with Disabilities Education Improvement Act (IDEIA) 2004, response to intervention (RTI) has now been given more attention regarding its role in the identification of students with and at-risk for LD (Marston, 2005). Longstanding concern about how LD is defined and identified, coupled with recent efforts in federal and state policy to eliminate IQ-achievement discrepancy as an LD marker, have led to serious public discussion about alternative identification methods (Fuchs, Mock, Morgan, & Young, 2003).

One of the issues critics of the discrepancy model for LD identification note is that the LD label is not just arbitrarily assigned, but unfairly withheld from children who are as needy and deserving but do not meet the criteria for LD (Fuchs et al., 2003). That is, students who struggle significantly in certain areas are equally in need of services but do not receive them because they do not meet the discrepancy criteria. Fuchs et al. (2003) argued that the IQ-achievement discrepancy approach also represents a "wait-to-fail" model since many students must perform poorly for years before their achievement scores are sufficiently below their IQ scores. Furthermore, the low achievement of many children labeled as LD may be due to poor teaching rather than disability, despite federal regulations stating that a child must have had appropriate learning experiences prior to labeling, and much of the necessary and costly data collection has as little to do with instruction as the label itself.

Response to Intervention

RTI is a problem-solving approach aimed, in part, at preventing unnecessary assignment to special education. It is described as providing help more quickly than traditional approaches to a greater number of struggling students, and also separates students with disabilities from those who perform poorly because of inadequate prior instruction (Marston, 2005). With RTI, before a low-performing child is designated for special education, he or she is offered intense, individualized academic intervention, which can occur through several (typically 3-4) tiers of intervention. The student's progress is evaluated and recorded to see if response to this intervention yields adequate academic growth. If the student is not responding, he or she will move to the next level (or tier) where more intensive services are given.

Research on Reading Problems, Remediation, and Prevention

Students with reading disabilities (RD) face enormous challenges learning to read, and many never reach a level of reading proficiency that allows them to achieve academically at a level commensurate with their intellectual abilities (Jenkins & O'Connor, 2001). Most researchers and practitioners agree that reading problems are more difficult to remediate than to prevent (Catts, Fey, Zhang, & Tomblin, 2001). Juel's (1988) longitudinal study that investigated reading development of 54 children from first to fourth grade found that the probability of a poor reader at the end of first grade remaining a poor reader at the end of fourth grade was .88.

Conversely, several studies have shown that early, intensive instruction can help prevent many at-risk children from developing reading difficulties. Lyon, Fletcher, Shaywitz, Shaywitz, Torgesen, Wood, et al. (2001) stated that the number of children typically identified as poor readers and served through special or compensatory education programs could be reduced by up to 70 percent through early identification and prevention programs. A study done by Foorman,

Francis, Fletcher, Schatschneider, and Mehta (1998) found that early instructional intervention significantly influenced the development and outcomes of reading skills in first- and second-grade children at risk for reading failure, where children who were directly instructed in the alphabetic principle improved in word-reading significantly more than children who received less direct instruction or regular class instruction, with effect sizes ranging from .69 to 1.13. Torgesen's review of 5 studies showed that when an intervention was used with the bottom 18 percent of the student population and works with 70 percent of them, the number of at-risk children requiring services is reduced from 18 percent to 5.4 percent, with the expected incidence of reading disabilities reduced from 12-18 percent to 1.4-5.4 percent.

Previous intervention studies (Pinnell, 1989; Wasik & Slavin, 1993) have demonstrated that through early and intensive remediation, most children who are experiencing reading difficulties can acquire at least grade level reading skills. Byrne, Fielding-Barnsley, and Ashley (2000) found that children who had been trained in phoneme identity 6 years earlier in preschool were superior to untrained children on several reading tasks. Elbro and Petersen's (2004) longitudinal study on the long-term effects of phoneme awareness training in kindergarten found that the 35 at-risk children who were trained outperformed the 47 untrained at-risk students in different reading tasks in Grades 2, 3, and 7. Therefore, early identification and prevention of reading difficulties is essential. However, school district personnel tend not to identify these children until the middle elementary grades, where reading difficulties have most likely grown "stronger roots" and have possibly become more difficult to remediate (Jenkins & O'Connor, 2001).

Early Literacy Screening

The first purpose for early literacy assessment is screening children to determine which children are at-risk for experiencing reading difficulties so additional instruction or intervention can be provided to these students (Coyne, 2006). Screening instruments are administered early in the school year so that those at-risk students can be identified as early as possible. Thus, it is imperative that these screening instruments be as accurate as possible in identifying those students who are at-risk. Speece (2005) posed the important question of how we are to identify the children who need secondary interventions; that is, what instruments do we use and what criteria should be in place for identifying children who need to move to a phase of instruction different from what is received by most children in general education?

To answer this question, Coyne (2006) stated that an essential feature of useful screening assessments is their predictive power, or their ability to accurately and reliably identify those students who will be most likely to experience reading difficulties in the future based on their current performance on critical indicators. However, despite strong correlations between phonological skills and later reading acquisition, predicting which children will have a reading disability has proved problematic (Jenkins & O'Connor, 2001). That is, measures that have strong correlations with reading often fail as effective screening tasks, yielding many classification errors (Speece, 2005). Overclassification often occurs, which refers to the screening procedure identifying children who do not experience reading problems (false positives). Some researchers of early literacy screening instruments were able to identify most of the students in their study who were at-risk for reading difficulties; however, because the criteria was set so that no child was missed, false positives ranged from 47-70% (Speece). Errors

of both under-prediction and over-prediction have made accurate early reading identification of students with reading disabilities difficult (Jenkins & O'Connor).

What are the ramifications of false positives and negatives in early literacy screening? Speece (2005) stated that if we are to accept as reasonable a 50% false positive rate, it means that we are comfortable identifying twice the number of children for secondary interventions than actually need them. This over-identification of students essentially consumes limited educational resources that are, in many school districts, already stretched very thin (Speece). When students are identified as at-risk for reading failure, but are actually not at-risk, resources such as personnel, money, and time, which are essential for effective education, are needlessly consumed. Furthermore, because schools have limited resources for students that are identified and moved into the second tier of intervention, the students who really need this instruction/intervention may not get it as intensively as they should when false positive students essentially "water down" this instruction. Finally, over-identification provides a skewed validation of interventions (Jenkins & O'Connor, 2001); that is, the false positives will show an "increase" in reading skills, when in actuality, the reading difficulties never existed. Thus, interventions in place for "at-risk" students will seem successful and receive skewed validation because the students seemingly caught up with their same-aged peers.

Because of these several factors, a system that can more accurately identify reading skill deficits in those "at-risk" students needs to be utilized. Other factors also need to be considered when developing an assessment system. For instance, cognitive development occurs rapidly during the younger stages of children's lives, and this development can vary for children of the same age. Therefore, any performance assessment measure needs to take into account the cognitive development of children at different ages in order to be a valid assessment of abilities.

Cognitive Development and Performance Assessments

Much research has been done regarding assessment and cognitive development of young children. Shepard, Kagan, and Wurtz (1998) stated that the younger a child is, the more difficult it is to obtain reliable and valid assessment, and that those involved in the assessment of children must be aware of the cognitive development of students in order to plan and implement effective and appropriate assessment. Assessing young children accurately is much more difficult than assessing older children and adults because of the nature of early learning and because early language skills needed to participate in formal assessments are still developing (Shepard et al.). Students in the earlier years of life are experiencing rates of physical, motor, linguistic, and cognitive development that outspace growth rates of all other stages of life. Doak and Chapman (1994) stated that informal as well as formal assessment must utilize valid and reliable instruments that convey accurate and useful information to students, teachers and parents. However, these informal and formal instruments are often fallible, and children vary greatly in their responses from one day to the next or in different contexts (Shepard et al.).

The unreliability of assessment instruments these authors speak of may result from children not having the experience to understand what the goals of formal testing are, thus making testing interactions very difficult or impossible to structure appropriately (Shepard et al., 1998). Children who are screened at the beginning of kindergarten have not had experience in formal testing procedures because they are just entering the school system, and therefore may not know what is expected of them. Also, kindergartners may not possess the cognitive skills it takes to perform well or even understand what they are to do on certain tasks. Thus, assessment at this level may reflect the child's developmental level rather than what the child knows (Doak & Chapman, 1994).

Cognitive development determines the amount of academic structure and direction needed by children as well as the type of academic activity students will find challenging instead of frustrating (Doak & Chapman, 1994). The variation in levels of cognitive development in children, therefore, requires that development of an assessment method or instrument that considers and accommodates for these developmental differences. For example, Wilson (1989) discussed the question of how memory and rehearsal or control processes interact in children. That is, efficiency of certain rehearsal processes allows some children to hold information and receive new information simultaneously, whereas other children are able to hold and process less information. This could become a problem when assessing children as young as kindergarten, especially when assessment measures require them to retain several different instructions at once and later remember these instructions when performing. Furthermore, Wilson provided evidence that control processes can be confounded with achievement effects. For example, processing and perceptual speed increase with age, so that tasks requiring more memory storage and retrieval may be very difficult for younger children but not be as difficult for older children on the same or comparable tasks.

Wilson (1989) argued that test developers often introduce other processing demands that are distinct from those intended, and that differentiating the new processing demands from the intended knowledge or processes required to solve the problem is difficult. Even an item or task's intended demands for certain information processes may unintentionally require the use of other information processes. Research in this area typically shows that children aged six years and below perform significantly differently than those aged seven years and older (Wilson), which would, in turn, have significant implications for the performance of kindergartners on tasks requiring processes involving memory storage and recall, information processing, and

rehearsal. Therefore, as Wilson suggested, test items or tasks should be developed with deliberate attention to the effect of memory and information processes required for their solution, and that test developers should state specifically what demands are required for each process intended for an item

Another issue that may result in the error variance associated with responses on assessment measures similar to the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) involves children's ability to make inferences. Like many cognitive and informational processes, the ability to make inferences develops and increases as children grow older, and more significant increases occur during the younger years of life. Thompson and Myers' (1985) found that four-year-old children are significantly less capable of making several different types of inferences than are seven-year-olds. Specifically, younger children have greater difficulty in making inferences based on what has been read. When read the (DIBELS) directions on the different probes, children are often asked to make inferences that they may not be capable of making. For instance, on the LNF probe, the directions do not specifically state that the children will be timed, that they need to go as quickly as they can, or that they need to continue after finishing a row of letters, until the examiner says, "Stop." Though this may be seen as an infrequent occurrence or possibly immaterial, the fact that some students may lack the ability to make such inferences could result in lowered scores, and consequently, a false positive at-risk identification for reading failure.

Dynamic Indicators of Basic Early Literacy Skills (DIBELS)

Within the RTI model, early literacy screening is typically the first phase in identifying those students who need more than what they are receiving in the regular classroom. There are several different instruments in use today for identifying at-risk students. One of the most

widely used instruments is the DIBELS. The DIBELS are a set of standardized, individually administered measures of early literacy development designed to be short, one-minute fluency measures used to regularly monitor the development of pre-reading and early reading skills ("Official DIBELS," 2006). According to Hintze, Ryan, and Stoner (2003), the DIBELS can be used in schools, especially with kindergarten and first-grade children, to address issues such as if (a) children are at-risk for reading difficulty because of inadequate phonological awareness skills, (b) children need additional instruction in phonological awareness skills, (c) current instruction is effective in increasing phonological awareness skills is necessary, and (d) a child has developed phonological awareness skills to a degree that is no longer indicative of difficulty learning to read. Langdon (2004) stated that there are no more powerful or meaningful measurements than DIBELS, and there is no more powerful tool for student and teacher accountability.

Research on DIBELS

In a prevention-oriented assessment and intervention system, foundational skills such as phonological awareness and alphabetic understanding can be assessed as early as the beginning of kindergarten and monitored over time as instruction changes and children's reading skills develop (Good, Simmons, & Kame'enui, 2001). According to a study done by Elliot, Lee, and Tollefson (2001), alternate forms reliability coefficients for the DIBELS kindergarten probes (with the exception of Initial Phoneme Ability [IPA]) ranged from .80 to .91. All reliability estimates (interrater, test-retest, and alternate forms reliability) were .80 or higher.

Such an instrument should show high validity and reliability with regard to the purposes of screening, assessment, and progress monitoring. Hintze et al. (2003) argued that if DIBELS is to be used for such educational purposes as resource allocation, placement, or identification, the

measures should be able to differentiate accurately between children who have not yet acquired such skills and those who have. Some research on the DIBELS has indicated that it is a valid and reliable instrument in terms of technical adequacy, identifying those students who are not making progress in acquiring early literacy skills and evaluating the effectiveness of interventions for individual students (Elliot, Lee, & Tollefson, 2001; Kaminski & Good, 1996). However, the use of suggested cut scores for identifying at-risk students in Hintze et al.'s study resulted in a large number of false positives (approximately 50%) on two of the DIBELS tasks. These authors suggest that using DIBELS (and these cut scores) could result in unnecessary allocation of resources to children, and children being inaccurately identified as "at-risk" for early reading problems.

If this is truly the case, and possibly 70% of students identified as at-risk are actually false positives (Speece, 2005), then the widespread popularity and use of the DIBELS is concerning. Manzo (2005) stated that teachers in Reading First (developed by the National Reading Panel and adopted under the No Child Left Behind Act of 2002) schools in more than 40 states now use the DIBELS to screen kindergarten through third grade students for potential reading problems, with some states adopting the assessments for all schools to use regularly. Within the Reading First schools, most of the 4800 schools use the DIBELS. This does not account for the numerous schools that are not Reading First schools and are utilizing the DIBELS for screening and assessment purposes. Although the total number of schools that are using the DIBELS is unknown, more than 8,200 schools in 2,600 districts have signed up for the associated data-management system offered by the University of Oregon at a cost of $1 per student (Manzo).

If the numbers of false positives are at such a high rate and are draining much needed resources, what is the solution to reducing this rate? Hintze et al. (2003) suggested using the DIBELS only for screening purposes, where all positively identified cases can be reassessed with instruments of higher specificity in order to identify false positives within the originally screened sample. This would result in basically screening children once to identify at-risk students, then screening those at-risk students to see which students actually are at-risk. This seems time-consuming and inefficient, especially considering how much time is already needed for the screening and assessment of early literacy skills in each child. Another possible solution would be to lower the cut scores for qualifying a student as at-risk; however, the obvious implications of doing this would be possibly raising the number of false negatives, or those students who are actually at-risk, but are not identified. So, are there any solutions to this problem that will not lead to other difficulties? An alternative that has not been addressed in the research on the DIBELS regarding screening for at-risk students is test-wiseness.

Test-wiseness

Many students, especially those in the younger stages of life, are unable to comprehend and implement directions that are given to them during test administrations or to answer questions in an effective manner, which results in invalid indicators of capability or achievement (Chance, 1992). Students will differ in their ability to perform on assessment measures, and this difference may be due to developmental level or prior experience with testing or situations similar to testing, and not to actual ability. One way to possibly control for this variance is through test-wiseness (TW) training. TW, which has also been called test-taking strategies/skills, test sophistication, test-taking orientation, and test-wisdom (Ying, 2001), refers to an individual's capacity to utilize characteristics and formats of a test or test-taking situation

to score well (Fuchs, Fuchs, Karns, Hamlett, Dutka, & Katzaroff, 2000). TW is, therefore, independent of an examinee's knowledge of what is being assessed, and these two characteristics need to be disentangled in order to evaluate students and instruction (Chance, 1992). The wide variation in cognitive development found among younger-aged children regarding comprehension and implementation of directions as well as the ability to answer questions effectively may be controlled for by providing students with TW-training prior to assessment. Thus, TW-training represents an attempt to "assure that all examinees possess a relatively equal level of the test-wise trait, so that the presence of test-wiseness cues will not differentially reward high test-wise examinees while concurrently penalizing individuals low in test-wiseness," (Sarnacki, 1979, p. 267).

Fuchs et al. (2000) stated the need for students in the primary levels to acquire specific forms of TW to be able to meet the demands of varying assessment formats and the tasks required of them. With the increasing reliance on performance measures like the DIBELS, this need takes on added importance because such measures differ dramatically from traditional achievement tests. Fuchs et al. stated that users of measures like the DIBELS may incorrectly infer that students lack the knowledge, skills, and strategies assessed, when it is actually a failure to demonstrate their competence in an unfamiliar assessment situation. Measurement experts agree that test results are invalid unless all examinees are sufficiently experienced and familiar with the testing situation to demonstrate their competence (Fuchs et al., 2000). Sarnacki (1979) describes one way in which children with lower levels of TW are at a disadvantage:

> The deductive test-taker is able to make use of their test-wiseness to improve his or her test score. Obviously, individual differences in test-wiseness determine how effectively one profits from... cues. However, the mere presence of such

> cues allows these individual differences to manifest themselves, thereby penalizing persons low in test-wiseness, and also affecting test reliability and validity. (p. 259)

Fuchs et al. (2000) provided an explanation of TW-training that involves brief orientation sessions, which allow students to demonstrate the competence they possess. The authors believe that by providing the brief TW-training, the validity of relatively unfamiliar assessments such as the DIBELS may be enhanced, which in turn provides the basis for more accurate generalizations regarding student competency and instructional effectiveness as well as permitting sounder decisions concerning the high-stakes consequences of referral and placement.

In Fuchs et al. (2000) study, TW-training showed dramatic effects on examinees' ability to demonstrate their knowledge and competence and enhanced the validity of students' scores in terms of their relation to another measure of competence. Specifically, scores on the performance assessments administered in the study increased dramatically when students had received the TW training, with effect sizes ranging from .54 to .75 on the different performance assessment components. In Roznowski and Bassett's (1992) study of TW training for flawed item types, results indicated that the TW training group showed significantly higher scores than those in either of the control groups (no effect sizes were given). Several of the studies in Bangert-Drowns, Kulik, and Kulik's (1983) meta-analysis of the findings on the effectiveness of coaching for achievement tests found effect sizes ranging from .45 to .85.

As mentioned earlier, another issue is the skewed validation of instruction or intervention following the identification of a student as at-risk. When initial testing is conducted without benefit of TW-training, because students are not test-wise, initial scores may be lower, and

higher scores on later performance assessments may not necessarily mean that students improved in that particular area or that the improvement was due to intervention. The classroom instruction received subsequent to initial testing may have incorporated a focus on TW, either implicitly or explicitly, and because of this, the higher scores may not be due to improvement in the student, just increased TW. An example of this is discussed in Fuchs et al. (2000), regarding Hambleton, Jaeger, Koretz, Linn, Millman and Phillips' 1995 study conducted on the measurement quality of the Kentucky Instructional Results Information System. In this study, although Kentucky children increased their scores over years on Kentucky's statewide performance assessment, they did not register commensurate improvements on the National Assessment of Educational Progress. Thus, it is difficult to say whether improvements occurred because of natural TW that may have occurred or because of this study's treatment conditions. This strengthens the proposition that at least some portion of the improvement reflected in current performance assessment accountability programs may reflect the kind of measurement error associated with TW.

Potential Problems with DIBELS

On the LNF probe, students are given an 8.5" x 11" sheet of paper with eleven rows of individual, randomly assigned letters, with ten letters in each row. The examiner has a sheet that has the same order of letters, on which he or she keeps track of incorrect responses. The student is to read the letters across each row as fast as possible for a period of one minute, after which the examiner tells the student to stop. The directions the examiner is to give are as follows:

> ***Here are some letters*** (point to the student probe). ***Tell me the names of as many letters as you can. When I say "begin," start here*** (point to first letter), ***and go across the page*** (point). ***Point to each letter and tell me the name of that letter.***

If you come to a letter you don't know I'll tell it to you. Put your finger on the first letter. Ready, begin.

This probe was chosen because it is the first probe administered to kindergartners upon entering school to assess their abilities naming letters of the alphabet. Because most students entering kindergarten have not been exposed to or trained in test-taking strategies, these probes may not provide an accurate assessment of their knowledge of the alphabet. For example, students may not know that they are being timed and therefore need to name letters as quickly as they can. Furthermore, although it says in the LNF directions to "start here (pointing to the first letter), and go across the page," students may not understand that they are to continue naming letters in that row or go on to the next row when completing the previous one. Students' different levels of TW may affect their scores, and those who are lower in TW may be falsely identified as at-risk simply because they are unfamiliar with what is expected of them.

The NWF measure is a standardized, individually administered test of the alphabetic principle - including letter-sound correspondence and of the ability to blend letters into words in which letters represent their most common sounds. The student is presented an 8.5" x 11" sheet of paper with randomly ordered vowel-consonant (VC) and consonant-vowel-consonant (CVC) nonsense words (e.g., sig, rav, ov) and asked to produce verbally the individual letter sound of each letter or verbally produce, or read, the whole nonsense word. The student is allowed one minute to produce as many letter-sounds as he/she can, and the final score is the number of letter-sounds produced correctly in one minute. Because the measure is fluency based, students receive a higher score if they are phonologically recoding the word and receive a lower score if they are providing letter sounds in isolation.

Look at this word (point to the first word on the practice probe). ***It's a make-believe word. Watch me read the word: /s/ /i/ /m/ "sim"*** (point to each letter then run your finger fast beneath the whole word). ***I can say the sounds of the letters, /s/ /i/ /m/*** (point to each letter), ***or I can read the whole word "sim"*** (run your finger fast beneath the whole word).

Your turn to read a make-believe word. Read this word the best you can (point to the word "lut"). ***Make sure you say any sounds you know.***

CORRECT RESPONSE: If the child responds "lut" or with all of the sounds, say	INCORRECT OR NO RESPONSE: If the child does not respond within 3 seconds or responds incorrectly, say
That's right. The sounds are /l/ /u/ /t/ or "lut"	***Remember, you can say the sounds or you can say the whole word. Watch me: the sounds are /l/ /u/ /t/*** (point to each letter) ***or "lut"*** (run your finger fast through the whole word). ***Lets try again. Read this word the best you can*** (point to the word "lut").

Place the student copy of the probe in front of the child.

Here are some more make-believe words (point to the student probe). ***Start here*** (point to the first word) ***and go across the*** *page* (point across the page). ***When I say, "begin", read the words the best you can. Point to each letter and tell me the sound or read the whole word. Read the words the best you can. Put your finger on the first word. Ready, begin.***

This probe was chosen for several reasons. First, the skills needed to perform well on this particular probe are more difficult and advanced than those needed for earlier-administered probes (i.e., LNF), which is why it is first administered in the middle of kindergarten and not at the beginning. Children will have had to develop more advanced skills (e.g., knowledge of the alphabetic principle, phonological recoding) in order to do well on this probe. This allows the researchers to examine skills that are more advanced and that are typically formed during the first part of kindergarten, if they are already not formed. Second, though information regarding the effect of the actual directions of the probe has not been examined, it seems that the directions for this probe can be somewhat confusing for children, especially children in kindergarten. For instance, the directions state that the child can identify the individual sounds or say the entire nonsense word. This assumes the child is aware that they can choose which method to use, even though it is stated in the directions, when this may not be the case. Furthermore, the directions go on to state, "Point to each letter and tell me the sound or read the whole word. Read the words the best you can." Again, these directions may confuse the child based on how they are structured; that is, it says first to either point to the letter and say the sound or read the whole word, but the very next sentence says *read the words* the best you can. TW-training would eliminate this error by giving the students the strategies and understanding of what they are to do on the test, allowing for the assessment of performance rather than test-taking knowledge.

Purpose of Study

The effects of brief TW-training on performance on the DIBELS kindergarten LNF and NWF probes and the influence of TW training on the performance of students on each of these probes were examined. The DIBELS probes are benchmark assessments, which are administered once at the beginning of the school year, once in the middle, and once at the end.

Students who are performing below a certain criterion level are considered at-risk for reading failure, and subsequent steps are taken to remediate the difficulties the student is experiencing and give him or her the skills that are "typical" of his or her same-aged peers.

The purpose of this study was to examine if children benefit from TW-training on the DIBELS. That is, were they be able to better answer the questions asked of them after they were exposed to specific testing situations prior to actual testing? The study's hypothesis was that those students exposed to testing directions, procedures, and situations prior to the actual testing would achieve significantly higher test scores than those who were administered the DIBELS in its traditional fashion.

Importance of Study

This study is potentially important for two reasons. First, the DIBELS assessment measure is being used more and more in school districts to identify students who might be at risk for reading failure. Second, if high numbers of false positives are being identified as at-risk for reading failure based on the results from these probes, then time, money, and additional resources are needlessly consumed and those who truly need the intervention are either not receiving it as intensively as is needed, or possibly at all. This is extremely important because the funds and other resources that are sometimes already scarce for some school districts will be unnecessarily depleted.

Method

Participants

Participants in this study consisted of approximately 60 kindergarten students in three different classes within a suburban Midwestern school district. Because the two parts of the study occurred at two different times during the school year, there was subject attrition resulting

in the loss of 3 subjects for the NWF part of the study. The district contains 13 K-5 elementary schools with approximately 80% of the students Caucasian, 6% African-American, 8% Hispanic, 4% Asian/Pacific Islander, 0.1% Native American, and 2% Multi-racial.

A power analysis indicated that at least 26 students should be in each group to detect large effects that are statistically significant. Large effects were expected based on previous research investigating the impact of TW training when other types of assessment have been used. Selection of participants was based on availability and willingness of the teacher as well as that of the students. The participant pool demonstrated near equal gender representation with 26 boys and 34 girls.

Participants in the study were randomly assigned into either the control group or the treatment group. Random assignment was completed using a random number generator designed by Daniels (2003; http://www.mdani.demon.co.uk/para/random.htm). The total number of students (62) was entered into the generator, which then produced an output of the numbers in a random order. Students were placed in alphabetical order and then assigned a number based on the random output (i.e. 1st student assigned the 1st number from output, 2nd student assigned the 2nd number from output, etc.). Once each student was assigned a number, those numbers were then placed in a box. As the numbers were randomly drawn from the box they were placed alternately in the control and treatment groups until all numbers (students) were placed in one of the groups.

Materials

There were two parts to this study, each utilizing a different probe from the DIBELS kindergarten probes. In the first part of the study, the Letter Naming Fluency (LNF) probe was used. For the LNF probe, students are presented with a page of upper- and lower-case letters

arranged in a random order and are asked to name as many letters as they can. Students are informed that if they do not know a letter, the name of that letter will be given to them, and they are to proceed to the next letter. Students are allowed one minute to produce as many letter names as they can, and the score is the number of letters named correctly in one minute.

According to the technical report provided by the DIBELS website ("Official DIBELS," n.d.) regarding the adequacy of the information obtained from the DIBELS subtests, the alternate-form reliability of LNF ranged from .86 to .92 in kindergarten (data were collected at 7 different points during one academic year). The concurrent validity of LNF with the Woodcock-Johnson Psycho-Educational Battery-Revised Reading Cluster standard score is .64 to .76 in kindergarten (data were collected at 8 different points during one academic year). The predictive validity of kindergarten LNF with first-grade Woodcock-Johnson Psycho-Educational Battery-Revised Reading Cluster standard score ranged from .44 to .69 (data were collected at 6 different points during one academic year).

In the second part of the study, the Nonsense Word Fluency (NWF) probes were used. For the NWF probe, the student is presented an 8.5" x 11" sheet of paper with randomly ordered VC and CVC nonsense words (e.g., sig, rav, ov) and asked to produce verbally the individual letter sound of each letter or verbally produce, or read, the whole nonsense word. For example, if the stimulus word is "vaj" the student could say /v/ /a/ /j/ or say the word /vaj/ to obtain a total of three letter sounds correct. The student is allowed 1 minute to produce as many letter-sounds as he/she can, and the final score is the number of letter-sounds produced correctly in one minute. Because the measure is fluency based, students receive a higher score if they are phonologically recoding the word and receive a lower score if they are providing letter sounds in

isolation. The NWF measure also takes about 2 minutes to administer and has over 20 alternate forms for monitoring progress.

According to the technical report provided by the DIBELS website regarding the adequacy of the information obtained from the DIBELS subtests, the alternate-forms reliability for NWF in January of first grade ranged from .67 to .88. The concurrent validity of DIBELS NWF with the Woodcock-Johnson Psycho-Educational Battery-Revised Reading Cluster score ranged from .35 to .59. The predictive validity of DIBELS NWF in January of first grade with (a) CBM ORF in May of first grade ranged from .68 to .82, (b) CBM ORF in May of second grade ranged from .63 to .85, and (c) Woodcock-Johnson Psycho-Educational Battery Total Reading Cluster score ranged from .52 to .77.

Procedures

Testing began in the fall of the 2006-2007 school year. Because the school already utilized the DIBELS for benchmarking and progress monitoring, informed consent for participation was not obtained. Prior to beginning the study, the researcher set up a meeting where the study was presented to those teachers whose classes were involved. The presentation outlined the purpose of the study, its importance, the procedures and materials needed, the time it would take to conduct, and the requirements of participants (both teachers and students). Also, two training sessions, each approximately 30 minutes long, were conducted with examiners, where specific procedures were practiced until all examiners were able to administer the training in the same way.

The first part of the study involving the LNF probes began approximately 3 weeks after the first day of school in September of 2006. The second part of the study involving the NWF

probes began approximately 3 weeks after the students returned from winter break in January of 2007.

Testing took place within the participating school in an unused classroom that was quiet and comfortable for the students. For each phase of each part of the study, students were taken out of the classroom and worked with on an individual basis in the available work space. Once testing with one student was completed, the student returned to his or her classroom and another student was then taken to the available work space and testing commenced.

For both parts of the study (LNF and NWF), students were randomly assigned to one of two groups. Each part of the study was broken into three different phases. Procedures for both the LNF and NWF parts were conducted in exactly the same manner, with the only differences being the script for each and the time of year.

LNF (beginning of kindergarten – within 3 weeks)
Phase 1 Phase 2 Phase 3

NWF (upon returning from winter break – within 3 weeks)
Phase 1 Phase 2 Phase 3

In Phase 1 (P1), the DIBELS kindergarten benchmark probe 1 was used. The participating school already used these probes for benchmarking data, therefore the data obtained were used as the "pretest," or P1. P1 occurred approximately 3 weeks after the beginning of the 2006-2007 school year. Testing with individual students in P1 took approximately five minutes.

When administering the probe in P1, the examiner administered the probe exactly how it is designed, using the directions provided in the DIBELS examiner booklet. This was done for both groups; therefore, administration of the probes in P1 was exactly the same for both the control and the treatment groups.

The second phase (P2) occurred approximately 3 weeks after the end of P1. It was felt this time was necessary in order to prevent any natural TW that may occur within the control group due to the administration of the probes in P1. Fuchs et al. (2000) discuss the "natural test-wiseness" that occurs through continued experience with testing throughout school years. If the examiner simply administered one probe right after the other, the students may develop, however major or minor, some degree of TW that would affect their performance on the P2 probe. Thus, by waiting three weeks to administer the P2 probe, any TW that the students may have gained and been able to utilize likely dissipated.

After the three weeks, the examiner conducted TW training with the treatment group using a script (provided in Appendix A). The scripts for both parts (LNF and NWF) were developed for this study and modeled after the script used in Fuchs et al. (2000). For the LNF part of the study, the training made the students aware (a) that they will be timed, (b) that they need to go as fast as they can, (c) that they need to continue until the examiner says, "Stop," (d) that they need to continue to the next line after finishing the previous line, (e) that they will be naming both capital and lowercase letters, (f) of how they will be scored, and (g) of what happens if they miss or don't know a letter. Training required students to participate and provide answers to questions that aid in developing TW. Following the training, students were given a quick review of the multiple aspects of the TW training. This review entailed the examiner stating the points that were covered throughout the training, beginning with the word, "Remember," followed by the TW training points (see LNF script in Appendix A).

For the NWF part of the study, the training made the students aware (a) that they will be timed, (b) that they need to go as fast as they can, (c) that they need to continue until the examiner says, "Stop," (d) that they need to continue to the next line after finishing the previous

line, (e) that they can *either* sound each sound out within the nonsense word *or* read the whole word, (f) of how they will be scored, and (g) of what happens if they miss or don't know a letter sound or word. Again, training required students to participate and provide answers to questions that aid in developing TW. Following the training, students were given a quick review of the multiple aspects of the TW training, which entailed the examiner stating the points that were covered throughout the training, beginning with the word, "Remember," followed by the TW training points (see NWF script in Appendix B).

Following the review at the end of each script, the students were given a practice test. During the practice test, a prompt was given if the student did not show TW. For example, if the student stopped after naming a letter, or if a student paused at the end of a row, the examiner would say, "Keep going until I say stop." Following this practice test, the examiner had one more review of the training in the form of a question and answer session that covered all aspects of the TW training for that particular probe (either LNF or NWF).

During P2 for both parts of the study (LNF and NWF), the control group was administered a practice test, which was given to control for the possible practice effects the TW training groups may have experienced due to the practice test following the training.

Immediately following the practice tests, both the control and treatment groups were then administered a third, alternate-form probe, which was Phase 3 (P3). When administering the third probe, the examiner administered the probe exactly how it was designed, using the directions provided on the DIBELS examiner booklet. This was done for both groups; therefore, administration of the probes in P3 was exactly the same for both the control and the treatment group.

LNF (beginning of kindergarten – within 3 weeks)

	Phase 1		Phase 2	Phase 3
Control	Form A	3 weeks	Form B	Form C
Treatment	Form A	3 weeks	TW training/Form B	Form C

NWF (upon returning from winter break – within 2 weeks)

	Phase 1		Phase 2	Phase 3
Control	Form A	2 weeks	Form B	Form C
Treatment	Form A	2 weeks	TW training/Form B	Form C

Testing in treatment P2 and P3 (administered one after the other) of the LNF section of this study took approximately 8-10 minutes per student. Because the script for NWF was longer and the tasks are more difficult than on the LNF, P2 and P3 (administered one after the other) took approximately 12-15 minutes per student.

Testing in control P2 and P3 (administered one after the other) of both the LNF and NWF sections of this study took approximately 5 minutes per student. Because the control group is simply being administered two alternate form probes for both LNF and NWF, P2 and P3 did not take as long as it did for the treatment groups.

Data Analysis

Table 1 lists the pretest, posttest, and gain score *M*s and *SD*s of the experimental and control group's DIBELS LNF and NWF scores.

Table 1

Pretest, Posttest, and Gain Score Means and Standard Deviations by Group

	Test-wiseness Group	**Control Group**
LNF Pretest	26.57 (16.34)	24.30 (14.40)
LNF Posttest	34.20 (18.88)	30.20 (15.73)
LNF Gain	7.63 (10.08)	5.90 (6.11)
NWF Pretest	40.83 (17.39)	38.59 (29.43)
NWF Posttest	50.17 (26.25)	41.41 (30.07)
NWF Gain	9.33 (17.32)	3.03 (7.94)

Letter Naming Fluency

A t-test using pretest LNF scores was conducted to determine whether the groups differed at the outset of the study on their letter naming fluency skill. Results indicated the groups possessed equivalent letter naming fluency skill at pretest (t (58) = .57, p = .57).

A t-test using posttest LNF scores was conducted to determine the impact of TW training. Results indicated no statistically significant differences between the groups at posttest (t (58) = .89, p = .38).

Because of the study's small sample size a t-test on gain scores was also conducted, despite no statistically significant differences at pretest. These results indicated the groups made similar gains from pretest to posttest (t (58) = .81, p = .42).

Each group was evaluated individually to determine whether LNF posttest scores were significantly higher than LNF pretest scores. Both the experimental group (t (29) = 4.15, $p < .01$,

d = .43) and the control group (t (29) = 5.29, p < .01, d = .39) showed statistically significant gains from pretest to posttest.

Nonsense Word Fluency

A t-test using pretest NWF scores was conducted to determine whether the groups differed at the outset of the study on the nonsense word reading skill. Results indicated no statistically significant differences at pretest (t (57) = .36, p = .72).

A t-test using posttest NWF scores was conducted to determine the impact of TW training. Results indicated no statistically significant differences between groups at posttest (t (57) = 1.19, p = .29).

Although the groups did not differ significantly at pretest, a t-test using gain scores was conducted to increase power. This analysis indicated that the differences in gain scores between groups approached statistical significance (t (57) = 1.79, p = .08, d = .47), with the gain scores for the experimental group higher than those for the control group.

Each group was evaluated individually to determine whether NWF posttest scores were significantly higher than NWF pretest scores. The difference between pretest and posttest scores approached statistical significance for the control group (t (28), p = .08, d = .10). Gains scores from pretest to posttest reached statistical significance for the experimental group (t (29) = 2.95, p < .01, d = .42).

Discussion

The effects of TW training on individual performance on early literacy measures were examined. Students were introduced to the structure of the early literacy measures, how they are scored, and the various expectations that are not stated explicitly. The purpose of this was to allow students to demonstrate competency they already possessed rather than increase students'

early literacy skills. Similar to Fuchs et al.'s (2000) study, the purpose of the TW training in this study was to reduce any construct-irrelevant variance associated with test-wiseness. All students took three alternate form probes, with the treatment group receiving TW training prior to the second "practice" probe, while the control group simply received the second "practice" probe. For the LNF part of the study, results indicated that there was no statistically significant difference between the treatment and control groups. For the NWF part of the study, results also indicated that there was no statistically significant difference between the treatment and control groups; however, the differences in gain scores between groups approached statistical significance. Therefore, had the sample been even slightly larger, a statistically significant difference would most likely have been seen between the treatment and control groups. Furthermore, when looking at gain scores from pretest to posttest for both groups, the experimental group reached statistical significance with a medium effect size of .42 *SD*, whereas the control group did not.

Table 2 shows the largest individual subject gains from pretest to posttest within the NWF treatment group. Investigating the individual cases within the treatment group where students received TW training revealed some significant improvements. When using DIBELS Benchmark Goals and Indicators of Risk for Nonsense Word Fluency, some students significantly improved their scores following the TW training and one student even moved from the "at-risk" range to the "normal" range.

Table 2

Individual gains from pretest to posttest within the NWF treatment group

Student	Pretest	Posttest
Student 1	5*	15
Student 2	14	35
Student 3	19	48
Student 4	25	50
Student 5	39	59
Student 6	40	63
Student 7	42	71
Student 8	55	77
Student 9	74	108
Student 10	80	144

*indicates scores in the "at-risk" range

Though most of the students were not in the "at-risk" range, it is important to look at the improvement made after the TW training. Since there were only two weeks between the pretest and posttest, it is likely that classroom instruction had not played a significant role in the improvement of students' scores within the treatment group. Furthermore, it is also likely that the students had not developed the decoding skills within those two weeks to allow them to improve their scores so drastically.

It is also important to look at the percentage of students who fell within the some risk to at-risk range during pretest of both groups, but specifically the treatment group. Table 3 shows the percentages of students overall and within each group that fell within the average, some risk,

and at risk ranges. Overall, only 7% (4 out of 59) of students in both groups fell within the some risk to at risk range at pretest according to DIBELS NWF Winter cut-scores. Within the treatment group, only 1 student fell within the some risk range at pretest (approximately 3%), while only 3 students fell within the some risk to at risk range at pretest (approximately 10%) in the control group. This may be important in that large differences between groups might not have been found because the majority of subjects in the sample population possessed adequate to advanced DIBELS-related skills. Perhaps a larger difference between groups may have been seen had more children scored in the at-risk range at pretest.

Table 3

*Percentages of subjects within the average, some risk, and at risk ranges at pretest**

	Percent in Average range	Percent in "Some Risk" Range	Percent in "At Risk" Range
Overall	93%	5%	2%
Treatment	97%	3%	0%
Control	90%	7%	3%

* Percentages based on percentile ranks provided by Good et al. (2002)

The majority of the students in this sample scored within the average range according to DIBELS Benchmark Goals and Indicators of Risk (approximately 92% of the total sample population). This was expected, as students within this particular school typically score higher on local and state assessment measures in the area of reading. However, even though the majority of students in this sample are well within the average range, the implications of this study's findings are important. First, there were several students who showed drastic improvements from pretest to posttest, even if they were not in the at risk range.

Second, students who score within the average to above average range on the DIBELS Benchmark Goals and Indicators of Risk are considered to have adequate early literacy skills, meaning they possess those basic early decoding skills to do well on this particular test. If students possess the skills to perform in the average to above average range on the pretest, then there should not be a drastic change from pretest to posttest, given there is no instruction in between that focuses specifically on this test or similar tasks. Yet, when looking at many of the individual scores, this drastic change exists from pretest to posttest, indicating some degree of test-wiseness.

There are several limitations to this study that should be addressed in future research. First, availability of subjects was limited to the researcher, as the study was conducted with subjects within one school. If possible, not only should the sample size be significantly larger, but should also use samples from several different schools and possibly schools not in the same district or schools that have more struggling readers. This will most likely increase the effect size as well as the ability to generalize any findings for the effects of test-wiseness training.

Second, training of examiners who participated (administered the probes) in this study would go through more rigorous training in how to administer the scripts. Though training sessions for following the scripts were adequate for both the LNF and NWF parts of the study (training for both scripts involved two 30 minute training sessions), more detailed training sessions would most likely increase the probability of higher reliability in examiner administration of the scripts. Similarly, all examiners had been previously trained in administering the DIBELS probes that were used in this study, and there was a biannual "refresher" course given to these examiners because the particular school district in which this study was done utilized the DIBELS for benchmarking purposes. However, there was not a

training or "refresher" course done with examiners pertaining to this particular study to ensure that all examiners administered the probes in exactly the same fashion. Future research may want to include a training session for administering the probes regardless if the examiners are experienced in giving the probes.

The DIBELS and similar measures are being used more and more within schools to identify students who may be at-risk for reading failure. It is essential that these measures be as accurate as possible in identifying those students who may be at-risk for reading failure so that students who actually need additional support receive it and those who do not are not incorrectly identified, ultimately saving schools from misallocating valuable resources. Although the results from this study did not yield statistically significant results, a larger sample size probably would have shown statistical significance. Nonetheless, when looking at the drastic improvements made by many of the students who received TW training, it seems that there was some factor outside of classroom instruction that improved their performance from pretest to posttest that can most likely be attributed to the TW training.

References

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1983). Effects of coaching programs on achievement test performance. *Review of Educational Research, 53*, 571-585.

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

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech, and Hearing Services in Schools, 32*(1), 38-50.

Chance, S. L. (1992). *Utilizing test wiseness to improve test scores in reading for eighth grade students* (Report No. CS011222). Fort Lauderdale-Davie, Florida: Nova Southeastern University, Center for Advancement of Education. ERIC Document Reproduction Service No. ED355475)

Coyne, M. D., & Horn, B. A. (2006). Promoting beginning reading success through meaningful assessment of early literacy skills. *Psychology in the Schools, 43*(1), 33-43.

Doak, J. L., & Chapman, A. D. (1994). Educational reform in early elementary assessment. *Journal of Instructional Psychology, 21*(1), 8-14.

Elbro, C., & Petersen, D. K. (2004). Long-term effects of phoneme awareness and letter sound training: An intervention study with children at risk for dyslexia. *Journal of Educational Psychology, 96*, 660–670.

Elliot, J., Lee, S. W., & Tollefson, N. (2001). A reliability and validity study of the Dynamic Indicators of Basic Early Literacy Skills – Modified. *School Psychology Review, 30*(1), 33-49.

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology, 90*(1), 37-55.

Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C. L., Dutka, S., & Katzaroff, M. (2000). The importance of providing background information on the structure and scoring of performance assessments. *Applied Measurement in Education, 13*(1), 1-34.

Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research & Practice, 18*, 157-171.

Good, R. H., III, Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foudational reading skills for third-grade high-stakes outcomes. Scientific Studies of Reading, 5, 257-288.

Good, R. H., III, Wallin, J. U., Simmons, D. C., Kame'enui, E. J., & Kaminski, R. A. (2002). System-wide Percentile Ranks for DIBELS Benchmark Assessment (Technical Report No. 9). Eugene, OR: University of Oregon.

Hintze, J. M., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the Dynamic Indicators of Basic Early Literacy Skills and the Comprehensive Test of Phonological Awareness. *School Psychology Review, 32*, 541-556.

Jenkins, J., & O'Connor, R. (2001). *Early identification and intervention for young children with reading/learning disabilities. Executive Summary*. (Report No. EC308705). Washington, DC: Special Education Programs (Ed/OSERS). (ERIC Document Reproduction Service No. ED458757)

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80*, 437-447.

Kaminski, R. A., & Good, R. H., III. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review, 23*, 215-228.

Langdon, T. (2004). DIBELS: A teacher-friendly basic literacy accountability tool for the primary classroom. *Teaching Exceptional Children, 37*, 54-58.

Lyon, G. R., Fletcher, J. M., Shaywitz, S. E., Shaywitz, B. A., Torgesen, J. K., Wood, F. B., et al. (2001). Rethinking Learning Disabilities [Electronic version]. In C. E. Finn, Jr., A. J. Rotherman, & C. R. Hokanson, Jr. (Eds.), *Rethinking Special Education for a New Century* (chap. 12). Washington, DC: Thomas B. Fordham Foundation & Progressive Policy Institute.

Manzo, K. K. (2005). National clout of DIBELS test draws scrutiny. *Education Week, 25*, 1-12.

Marston, D. (2005). Tiers of Intervention in Responsiveness to Intervention: Prevention outcomes and learning disabilities identification patterns. *Journal of Learning Disabilities, 38*, 539-544.

Official DIBELS Home Page (n.d.). *DIBELS data system*. Retrieved February 21, 2006, from University of Oregon Center on Teaching and Learning Web site: http://dibels.uoregon.edu/index.php

Roznowski, M., & Bassett, J. (1992). *Training test-wiseness and flawed item types. Applied Measurement in Education, 5*(1), 35-48.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive domain. *Review of Educational Research, 49*, 252-279.

Shepard, L., Kagan, S. L., & Wurtz, E. (Eds.). (1998). *Principles and recommendations for early childhood assessments* (Report No. PS026296). Washington, DC: National Education Goals Panel. (ERIC Document Reproduction Service No. ED416033).

Speece, D. L. (2005). Hitting the moving target known as reading development: Some thoughts on screening children for secondary interventions. *Journal of Learning Disabilities, 38*, 487-493.

Thompson, J. G., & Myers, N. A. (1985). Inferences and recall at ages four and seven. *Child Development, 56*, 1134-1144.

Torgesen, J. K. (2000). Individual responses in response to early intervention in reading: The lingering problem of treatment registers. *Learning Disabilities Research & Practice, 15*, 55-64.

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

Wilson, V. L. (1989). Cognitive and developmental effects on item performance in intelligence and achievement tests for young children. *Journal of Educational Measurement, 26*, 103-119.

Ying, P. (2001). Effects of test-wiseness upon performance on the Test of English as a Foreign Language. (Doctoral dissertation, University of Alberta, 2001). *Dissertation Abstracts International Section A: Humanities and Social Sciences, 62*(5-A), 1724.

APPENDIX A

DIBELS LNF Test-wiseness Training Script

Directions for LNF probe

Here are some letters (point to the student probe). Tell me the names of as many letters as you can. When I say "begin", start here (point to first letter), and go across the page (point). Point to each letter and tell me the name of that letter. If you come to a letter you don't know I'll tell it to you. Put your finger on the first letter. Ready, begin.

Potential problems with DIBELS directions

It is not explicitly stated to students:

- that they need to go as fast as they can
- that they are timed
- how they are scored
- that they go until the examiner says stop
- that students need to continue to the next line after finishing the previous line
- the letters are both in capital and lowercase, which may need to be explained so the students are aware of what will be on the test

LNF Script

Examiner: **We took a test a few weeks ago in which you named letters; do you remember? You were asked to tell me the names of as many letters as you can, and the letters were all mixed up. Now, I want to tell you the best way to take this test *and* how it is scored. Then, you'll take another test just like it. Listen carefully.**

First, I will be timing you. You will have *1 minute* to name as many letters as you can.

How long do you get to take this test?

Students: 1 minute (If incorrect or no response say "You get 1 minute to name as many letters as you can").

Examiner: **Right! Once I say begin, you are to name as many letters as you can until I say stop. Keep naming letters until I say, "stop." That means, once you finish saying all the letters in one row** (using the probe administered in the phase 1, the examiner moves finger across the top row of the student sheet used for the first test), **go on to the next row without stopping** (move finger down to second row and across). **After I say "Begin," when should you stop naming letters?**

Students: When you say, "stop." (If incorrect or no response, say "You stop naming letters when I say stop").

Examiner: **Very good! The letters that you will be seeing will be in either lowercase, which are small letters** (tester points the "d" in the second row) **or capital, which are big letters** (tester points to the "N" in the second row). **This means you will have to name some lowercase letters and some capital letters. Will some letters be in lowercase?** (Wait for student response – say "Yes, some letters will be lowercase" if incorrect or no response). **Will some letters be in capital?** (Wait for student response – say "Yes, some letters will be capital" if incorrect or no response).

Examiner: **Great! Now, on this test, you are able to score points, and we want you to score *as many points as possible*. The way you get points is by correctly naming letters. That means for every letter you get right, you get a point! In order to get as many points as you can, you have to name as many letters as you can until I say "stop." How do you get points?**

Students: By naming letters (correctly) - If incorrect or no response, say "You get points by naming letters correctly".

Examiner: **Right! Now, if you don't know a letter, I will tell you what it is, then you move on to the next letter. What happens if you don't know a letter?**

Students: You tell me the letter (and I move on to the next letter).

Examiner: **Right! Now, I want you to do your best and get as many points as you can. To do this, I want you to read the letters as fast as you can so you can get as many points as you can. Do you understand?**

Students: (Acknowledges that he/she understands)

Examiner: **Great!**

OR

That's okay, let's go over it again.

The examiner will then say:

"Remember, you will get 1 minute, so keep going until I say stop and go as fast as you can. The letters will be in capital and lowercase and you will get points by correctly naming the letters. Remember, the more letters you name correctly, the more points you get. If you don't know a letter, I'll tell it to you and then you move on to the next letter. Do you understand?"

Examiner: **"Okay, let's practice."**

Examiner switches to Practice Probe 1

Administer the practice probe using the same directions provided for the LNF probe. Begin timing. Throughout the 1-minute probe, if the student does not show TW, prompt the student

accordingly. For example, if the student stops after naming a letter, or if a student pauses at the end of a row, the examiner says, "Keep going until I say stop."

Examiner: **Good Job! Now, we'll take another test where you will name letters. Let's go through how we should take the test and how it is scored one more time** [go through strategies and scoring procedures].

Examiner will ask (prompt if student does not respond):

- **How long do you get to take this test?** (*1 minute*)
- **After I say "Begin," when should you stop naming letters?** (*When you say Stop*)
- **Will some letters be in capital?** (*Yes*)
- **Will some letters be in lowercase?** (*Yes*)
- **How do you get points?** (*By naming letters correctly*)
- **What happens if you don't know a letter?** (You *tell me the letter and I move to the next letter*)
- **How fast do you go?** (*As fast as I can*)

APPENDIX B

DIBELS NWF Test-wiseness Training Script

Directions for NWF probe

Look at this word (point to the first word on the practice probe). It's a make-believe word. Watch me read the word: /s/ /i/ /m/ "sim" (point to each letter then run your finger fast beneath the whole word). I can say the sounds of the letters, /s/ /i/ /m/ (point to each letter), or I can read the whole word "sim" (run your finger fast beneath the whole word).

Your turn to read a make-believe word. Read this word the best you can (point to the word "lut"). Make sure you say any sounds you know.

CORRECT RESPONSE: If the child responds "lut" or with all of the sounds, say	INCORRECT OR NO RESPONSE: If the child does not respond within 3 seconds or responds incorrectly, say
That's right. The sounds are /l/ /u/ /t/ or "lut"	Remember, you can say the sounds or you can say the whole word. Watch me: the sounds are /l/ /u/ /t/ (point to each letter) or "lut" (run your finger fast through the whole word). Lets try again. Read this word the best you can (point to the word "lut").

Place the student copy of the probe in front of the child.

Here are some more make-believe words (point to the student probe). Start here (point to the first word) and go across the page (point across the page). When I say, "begin", read the words the best you can. Point to each letter and tell me the sound or read the whole word. Read the words the best you can. Put your finger on the first word. Ready, begin.

Potential problems with DIBELS directions

It is not explicitly stated that students:

- need to go as fast as they can

- are timed
- need to go until the examiner says stop
- need to continue to the next line after finishing the previous line
- or, how they are scored

Also, it is not exactly clear that the student can either say each sound in the nonsense word or read the whole word. The directions can be confusing to the child. For example, in the directions, it states that the child is to, "Point to each letter and tell me the sound or read the whole word," but then the very next sentence states "Read the *words* the best you can." These conflicting statements may confuse the student. Providing the students with TW training will allow them to understand that they can use either strategy and give them practice in using that strategy.

NWF Script

Examiner: **We took a test a few weeks ago in which you read make-believe words; do you remember? I want to tell you the best way to take this test and how it is scored. Listen carefully.**

First, I told you that you could point to each letter in a word and tell me the sound it makes *OR* you could read the whole word. That means if you think it would be easier for you to say what sound each letter makes, you can do it that way. Let's try saying the sounds of the letters – "s – i – m". Good! You try this one – "l – u – t". Good! If you think it would be easier for you to read the whole word, then you can do it that way. Let's try – "sim". Good! You try this one – "lut". Great!

Remember, you can do it whichever way is easiest for you, but *YOU ONLY HAVE TO DO IT ONE OF THOSE WAYS*. So, what is one way you can do this test?

Students: Point to each letter and say what sound it makes (or read the whole word).

Examiner: **Right. And what is the other way you can do this test?**

Students: Read the whole word (or point to each letter and say what sound it makes).

Examiner: **Great job! I will be timing you. You will have *1 minute* to read as many words as you can. How long do you get to take this test?**

Students: 1 minute.

Examiner: **Right! When I say begin, read as many words as you can until I say stop. Keep going until I say, *"STOP."* That means, once you finish all the words in one row** (move finger across top row of the student sheet used for the first test), **go on to the next row without stopping** (move finger down to second row and across). **After I say "Begin," when should you stop?**

Students: When you say, "stop." (If incorrect/no response, say "Stop naming letters when I say stop").

Examiner: **Great! On this test, you are able to score points, and we want you to score as many points as possible. To get as many points as you can, you have to correctly name as many letter sounds or correctly read as many words as you can until I say "stop". So, if you choose to tell me the sound each letter makes, how do you get points?**

Students: By correctly naming the sound each letter makes

Examiner: **Correct. If you choose to read the whole word, how do you get points?**

Students: By reading the word correctly (if incorrect or no response, say, "by reading the words correctly")

Examiner: **Right! If you don't know a sound or a word, I will tell you how to say it and you move on to the next sound or word. What happens if you don't know the sound a letter makes or a word?**

Students: You tell me the letter sound or word and I move on to the next word.

Examiner: **Right! I want you to do your best and get as many points as you can. To do this, go as fast as you can so you can get many points. Do you understand?**

Students: (Acknowledges that he/she understands)

Examiner: **Great!**

OR

That's okay, let's go over it again.

The examiner will then say:

"Remember, you will get 1 minute, so keep going until I say stop. Go as fast as you can. You can point to each letter and tell me the sound it makes *OR* you can read the whole word. Remember, the more sounds you name correctly, or the more words you read correctly, the more points you get. If you don't know a sound or a word, I'll tell it to you and you move on to the next one. Do you understand?"

Examiner: **"Okay, let's practice."**

Examiner switches to Practice Probe 1

Administer the practice probe using the same directions provided for the NWF probe. Begin timing. Throughout the 1-minute probe, if the student does not show TW, prompt the student accordingly.

Examiner: **Good Job! Now, we'll take another test just like that one. Let's go through how we should take the test and how it is scored one more time** [go through strategies and scoring procedures].

Examiner will ask:

- **How long do you get to take this test?** (*1 minute*)
- **After I say "Begin," when should you stop reading?** (*When you say Stop*)
- **What is one way you can take this test?** (*Point to each letter and say the sound it makes)*
- **What is one way you can take this test?** (*read the whole word*)
- **How do you get points?** (*By naming letter sounds correctly or by reading the word correctly*)
- **What happens if you don't know a letter sound or word?** (*You tell me the letter sound or word and I move to the next one*)
- **How fast do you go?** (*As fast as I can*)