

Assessing the Instructional Level for Mathematics: A Comparison of Methods.

by Matthew K. Burns , Amanda M. VanDerHeyden , Cynthia L. Jiban

Abstract. This study compared the mathematics performance of 434 second-, third-, fourth-, and fifth-grade students to previously reported fluency and accuracy criteria using three categories of performance (frustration, instructional, and mastery). Psychometric properties of the fluency and accuracy criteria were explored and new criteria for the instructional level were empirically derived. Two sets of mathematics probes were administered to the students and delayed alternate-form reliability coefficients were obtained from multiskill probes. Results suggested that fluency data were significantly more reliable than accuracy data. Slopes from the single-skill probes were used to categorize students as high responders, and the mean fluency scores from that group's multiskill probes were used to suggest an alternative instructional level of 14-31 digits correct per minute for second- and third-graders and 24-49 digits correct per minute for fourth- and fifth-graders. Implications for practice and suggestions for future research are discussed.

Recent empirical attention to academic intervention has focused primarily on reading (Badian, 1999; Daly & McCurdy, 2002). Research related to mathematics assessment and instruction has lagged comparably behind, yet the need for evidence in effective mathematics assessment and intervention is pressing because fewer than one third of fourth-grade students met or exceeded the proficiency standard on the 2003 mathematics test of the National Assessment of Educational Progress (Manzo & Galley, 2003).

Panels convened to recommend reform in mathematics assessment and instruction have emphasized, among other factors, the need for data that are useful to teachers in planning and delivering **math** instruction (National Council for Teachers of Mathematics, 2000). Moreover, assessment data can be used for progress monitoring and to facilitate attainment of desired outcomes (Algozzine, Ysseldyke, & Elliott, 1997). Thus, instructionally relevant assessment data are critical to effective mathematics instruction and intervention.

Academic difficulties can result from a mismatch between student skill and the curriculum or instructional material (Daly, Martens, Kilmer, & Massie, 1996; Daly, Witt, Martens, & Dool, 1997; Enggren & Kovaleski, 1996; Gravois & Gickling, 2002). Instructional material that is too difficult results in student frustration, and material that is not challenging enough, or is too easy, results in student boredom. The "window of learning" (Tucker, 1985, p. 201) between boredom and frustration is called the "instructional level" and occurs when instructional materials provide an appropriate level of challenge. Research has consistently found that providing appropriately challenging teaching material, at the instructional level, has led to improved student outcomes for reading (Burns, 2002; Gickling & Rosenfield, 1995; Shapiro, 1992; Shapiro & Ager, 1992)

and mathematics (Burns, 2002; Gickling, Shane, & Croskery, 1989).

An appropriate level of challenge (or instructional level) is one of the essential components of an effective learning environment (Ysseldyke & Christenson, 2002), but research has yet to adequately define an instructional match for mathematics. Gickling and Thompson (1985) suggested an accuracy approach in which mathematics assignments should contain 70-85% known items to represent an instructional level task. Deno and Mirkin (1977) suggested that the instructional level for mathematics be determined with fluency (i.e., accuracy plus speed) measures instead of accuracy data alone. They further estimated that 10-19 digits correct per minute (dc/min) would represent an instructional level for students in the 1st through 3rd grades, whereas 20-39 dc/min would equal an instructional level for children in the 4th through 12th grades. Meta-analytic research has found strong effects from studies using the accuracy criterion of 70-85% known items for mathematics, but also found strong effects for other proposed accuracy criteria such as 50% known and 90% known (Burns, 2004). Previous research also supported the fluency criteria to the extent that instructional adaptations based on fluency criteria were found to improve student learning (Daly & Martens, 1994; Daly et al., 1996; VanDerHeyden & Burns, 2005). It is important to note that the format of fluency and accuracy assessments are generally similar, with the difference being the scoring metric. Specifically, these tasks are typically timed. Hence, although one is scored with an accuracy metric, because it is a timed task, fluency affects the score. Similarly, the fluency measure is obtained through a timed task but is scored typically as responses correct per unit of time. Therefore, fluency also captures accuracy to some degree.

The use of fluency measures to determine optimal challenge for mathematics makes some intuitive sense, given that fluent computation is a goal for mathematics instruction (National Council of Teachers of Mathematics, 2000). However, no data were provided to demonstrate how the fluency standards suggested by Deno and Mirkin (1977) or the accuracy criteria proposed by Gickling and Thompson (1985) were derived and only limited data exist to support the validity of either approach. Deno and Mirkin (1977) stated that the criteria were established at a school that was part of the precision teaching program being conducted in Minnesota (S. L. Deno, personal communication, April 15, 2005), but the manual provided no other information about how the criteria were developed and provided no additional data. The lack of an empirically researched definition for the instructional level for mathematics could make interpretation of assessment data difficult, but does not detract from the need to assess it.

Further complicating interpretation of mathematics performance data is the relative paucity of data supporting basic technical properties of decisions made based upon mathematics performance. Assessment tools used for educational decision making must meet certain criteria with technical data for each purpose for which the assessment tool is used (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). Reliability of test data with which academic growth is measured should be estimated with alternate-form or test-retest methods using separately timed administrations (AERA et al., 1999). Reliability coefficients for fluency estimates of mathematics performance have been reported to exceed .90 for delayed administrations (Tindal, Germann, & Deno, 1983), concurrent administrations of alternate forms (Thurber, Shinn, & Smolkowski, 2002), internal consistency (Fuchs, Fuchs, & Hamlett, 1989),

and consistency between scorers (Hintze, Christ, & Keller, 2002; Tindal et al., 1983). Foegen and Deno (2001) found alternate-form and test-retest reliability coefficients for mathematics fluency scores that ranged from .67 to .84 for individual probes. However, several scholars have suggested specifically that a delayed alternate-form approach be used to estimate reliability of scores (Anastasi & Urbina, 1997; Murphy & David-shofer, 2001). The delayed alternate-form approach involves administering one form of the test, administering a second form after a test-retest interval of at least 2 weeks, and correlating the two scores (Salvia & Yssledyke, 2004). Further, no studies were found that examined the reliability of assessing the instructional level for mathematics using either fluency or accuracy.

Current standards for assessment also call for evidence of validity in assessment data uses based on response processes, internal structure, relations to other variables, and consequences of testing (AERA et al., 1999). Relationships between assessment data and variables external to the test can be useful sources of validity evidence because these relationships estimate the degree of consistency between assessment and the construct upon which data interpretations are based (AERA et al., 1999). Although several approaches exist for examining validity of data, criterion-related validity is the simplest method with which validity can be empirically estimated (Murphy & Davidshofer, 2001). Moreover, most methods of evaluating validity generally involve examining the relationship between performance during the assessment and during other observations involving the assessed behavior (Anastasi & Urbina, 1997).

Criterion validity of computational fluency-based progress-monitoring measures was assessed in various studies, with correlations between fluency measures and standardized mathematics tests typically falling in the moderate range (e.g., .30-.60). Previous studies examined criterion validity of a basic mathematics facts measure, wherein only single-digit fact families from all four operations were mixed in a single measure. Coefficients from those studies ranged from .35 to .56 for third- through fifth-grade students (Espin, Deno, Maruyama, & Cohen, 1989), .45 to .67 for fourth-grade students (Thurber et al., 2002), and .35 to .61 for sixth-grade students (Foegen & Deno, 2001). When the measure represented computational skills beyond simple mathematics facts, correlations with criterion mathematics tests were also moderate (0.26-0.65--Skiba, Magnusson, Marston, & Erickson, 1986; .38-.63--Thurber et al., 2002). None of these studies focused on use of these measures in determining an instructional level, but instead used the distribution of raw scores to examine criterion validity.

The importance of providing a match between student performance and instructional techniques employed (Ysseldyke & Christenson, 2002) and the apparent lack of data examining the assessment of the instructional level for mathematics suggests the need to empirically investigate the technical adequacy of decisions based upon instructional ranges in mathematics. Therefore, the current study examined the reliability and criterion validity of the fluency and accuracy criteria used to identify the instructional level for mathematics. Three research questions guided this study: Which approach to assessing the instructional level, fluency or accuracy, leads to stronger delayed alternate-form reliability coefficients? Which approach to assessing the instructional level, fluency or accuracy, correlates better with scores from a standardized commercially available test of mathematics? Could student growth data obtained during protocol-based **math** intervention be used to suggest potential criteria for an instructional level? The last question operated under the logic that the fastest rate of student growth (i.e., strongest

slope) in response to a standard intervention would occur for students for whom the task difficulty was at an instructional level.

Method

Participants

Four hundred and thirty-four students from an elementary school located in a suburban community in the southwestern region of the United States participated in this study. The elementary school included four classrooms per grade level, with approximately 25 students per class. All students in Grades 2-5 participated in the study (28% were second-grade students, 25% were third-grade students, 27% were fourth-grade students, and 20% were fifth-grade students). Sex was equally distributed (49% female). Seventy-four percent of students were described as Caucasian (not Hispanic), 17% were described as Hispanic or Latino, 6% were described as African American, 3% were described as Asian American, and 1% were described as Native American. Twenty-six students (6%) received special education services because of a diagnosis of learning disability, emotional and behavioral disorder, or mental retardation. School-wide data suggested that about 3% of the students were eligible for Title I services. Whereas 17% of the sample were children from a Hispanic background, only 2% were English-language learners, which may under-represent this population as compared to many school districts. Finally, 15% of children at this school received free or reduced-cost lunch, which again may underrepresent this group relative to other school districts.

Measures

Two types of mathematics probes were administered weekly as part of a problem-solving model using state instructional frameworks for **math** instruction. A single-skill probe was used to track response of children to a standard intervention and to determine when to increase task difficulty. Second, a mixed probe was administered each week to track retention of skills that had been previously mastered during the intervention.

All probes were computer generated on a single sheet of white paper with a place at the top for the student's name, date of probe, and title (e.g., Weekly Probe Basic Facts, Fourth Grade), and were administered based upon procedures described by Shinn (1989), with one difference. Shinn's (1989) instruction to place an "X" over unknown problems and continue on to the next problem was omitted and students were instructed, "Try to work each problem. Do not skip around." This alteration was made because with group administration it was impossible for adults to ensure that students were attempting to work on each problem as opposed to only attempting the easiest problems on the paper. If a student attempted only the easiest problems (e.g., just the addition problems on a mixed probe or only the easier problem types such as adding ones and twos and multiplying zeros, ones, and fives), the probe would not reflect a child's competence on the full range of skills represented on the probe. Probes were group administered (to the class as a whole) by the classroom teacher using scripted instructions. Students were allowed 2 min to complete as much of the probe as they could and were instructed to not omit any problems. The latter direction was monitored by the classroom teacher who administered the probe. The teacher was trained to walk around the classroom while students

were working and to prompt children who were "stuck" on a particular problem to move on to the next problem. This approach allowed the adult to ensure that children did not unduly skip around doing the easier problems, while at the same time ensuring that children did not spend the entire 2 min stuck on a single problem, unduly deflating the student's score.

Single-skill probe. The single-skill probe used to assess progress on the classwide intervention and to determine when to increase task difficulty differed between grades and item content. Second- and third-grade probes consisted of randomly selected problems with one skill presented in eight rows and five problems per row. Fourth- and fifth-grade probes also contained eight rows, but with seven problems in each. Items sampled in the probes were taken from the instructional sequence presented in the Appendix. Each skill was probed on a weekly basis until the class median score exceeded the grade-specific fluency criteria for an instructional level suggested by Deno and Mirkin (1977).

Mixed-skill probe. Students were administered a mixed probe of mathematics problems taken from the skills listed in the intervention skill sequence. Mixed probes were used because previous research on technical adequacy of fluency measures for mathematics used mixed probes (Foegen & Deno, 2001; Fuchs et al., 1989; Hintze et al., 2002; Tindal et al., 1983; Thurber et al., 2002).

Different mixed probes were used for different grades within the school. Second- and third-grade probes consisted of 40 randomly selected problems, eight rows with 5 problems per row. The second-grade probe consisted of a mix of addition and subtraction problems with two single-digit numbers or one single-digit and one double-digit number with answers to 20. Third-grade probes consisted of a mix of problems including addition and subtraction facts 0-20, addition and subtraction of two three-digit numbers without regrouping, multiplication facts 0-9, and division facts 0-9. Fourth- and fifth-grade probes consisted of 56 problems arranged in eight rows containing 7 problems each. All categories of items included in the third-grade probes were also included in the fourth- and fifth-grade probes with the following differences: Addition and subtraction problems were all double-digit and single-digit or double-digit and double-digit, both with and without regrouping. Multiplication and division problems sampled facts 0-12. For all fact problems on all probes, a fill-in-the-missing-number format was used so that occasionally the answer number was missing, but other times the student had to supply the divisor when given the dividend and the answer. Placement of the missing number (e.g., dividend, divisor, answer) was randomly selected then counterbalanced. The order of problem types within all probes was randomly determined, with an approximately equal number of problems sampling each problem type. The first probe used for the current study was administered in March 2004. Two weeks after the initial data collection session, a second set of data were obtained by administering a second mathematics probe to the children. The second probe was identical to the first in format, but contained different items.

Stanford Achievement Test. Fluency and accuracy scores were compared to performance on the Stanford Achievement Test (9th ed.; SAT-9; Harcourt Brace Educational Measurement, 1996). The SAT-9 was administered using standardized administration procedures in April of the school year as part of the state's accountability program. Item content for the SAT-9 was aligned with the National Assessment of Educational Progress and the Curriculum and Evaluation Standards

for School Mathematics described by the National Council for Teachers of Mathematics. The standardization sample for the SAT-9 included 450,000 students enrolled in the spring and fall of 1995 and was stratified by ethnicity, urbanicity, and socioeconomic status. Reliability coefficients (K-R20 and alternate form) have been reported in the .80-.90 range for the total **math** cluster and **math** multiple-choice portion of the test. Performance standards (i.e., criterion reference) have been described, but no reliability estimates have been reported. In addition, SAT-9 scores have been reported to correlate strongly with previous versions of the SAT and with the Otis-Lennon School Ability Test (Haladyna, 1998).

Instructional Setting

Mathematics instruction occurred in the regular classroom. Eighty-four percent of teachers were female. The greatest number of teachers had 4-6 years of teaching experience at the time of this study (31%). Nineteen percent of teachers had less than 1 year of teaching experience and 16% had more than 10 years of teaching experience. All teachers were certified in elementary or special education, and 25% had master's degrees.

The school where this study was conducted provided instruction according to the district calendar that specified when certain skills should be taught to cover all the skills listed in the state standards for each grade level. Second- and third-grade teachers used the Saxon mathematics curriculum and materials (Saxon Publishers, 2004) to facilitate mathematics instruction, whereas fourth- and fifth-grade teachers used Harcourt **Math** and materials (Harcourt Publishing, 2003).

Estimating Reliability of Probe Scores

Reliability coefficients for fluency and accuracy scores were correlated to address the first research question. Scores from mixed probes were converted to a digits correct per minute metric by counting the number of digits correctly answered (Shinn, 1989) and then dividing by 2 to convert the 2-min probe to a 1-min metric. Accuracy was determined by dividing the number of digits correctly answered by the total number of possible digits completed. For example, if a student completed 10 problems with two potential digits correct for each one, there would be a total of 20 possible digits correct. Thus, if this student correctly completed 18 of the digits from those 10 problems, the accuracy score would be 90%. Fluency and accuracy scores were correlated between the two probes with Pearson product moment correlation coefficient.

Fluency and accuracy data were then converted to categories of frustration, instructional, and mastery using the fluency categories outlined by Deno and Mirkin (1977) and the accuracy categories described by Gickling and Thompson (1985). The resulting category scores were correlated with Kendall's tau. Fluency was compared to accuracy by converting the Pearson product moment and Kendall's tau correlation coefficients to z with Fisher's transformation. All analyses were conducted using the grade groupings used by Deno and Mirkin (1977).

The second research question asked which approach to assessing the instructional level, fluency or accuracy, correlated with scores from a commercially available standardized test of mathematics. This was evaluated by correlating scores from the first mixed probe with age-based

standard scores obtained on the SAT-9 mathematics test using a Pearson product moment correlation. Categorical data from the mixed probes were also correlated with SAT-9 standard scores using Spearman's rho. The fluency and accuracy approaches to estimating the relationship between the probes and SAT-9 scores were examined by again comparing the aggregate coefficients using Fisher's z transformation.

Use of Intervention Data to Derive Instructional Ranges

Intervention occurred 4 days per week for all students following a particular sequence of skills and a standard intervention protocol as a supplement to regularly scheduled math instruction. Children were divided into tutoring pairs and a format of intervention was used similar to classwide peer tutoring (Greenwood, 1991) and peer-assisted learning strategies (Fuchs, Fuchs, Mathes, & Simmons, 1997; Fuchs, Fuchs, Phillips, Hamlett, & Karns, 1995). The key elements of the intervention included paired peer instructional-level skill practice with multiple massed opportunities to respond at a brisk pace and with immediate corrective feedback, independent timed practice for a score, slightly delayed corrective feedback, and progression to the next skill level contingent on meeting a mastery criterion on the lower level skill. A protocol for the intervention was developed and provided to teachers that specified each step of the intervention in observable terms.

A total of 19 integrity observations were performed from November 2003 to March 2004. A trained observer held a copy of the teacher's intervention script, which specified each observable step of the intervention that was to occur and placed a check mark next to all independently and correctly implemented steps. The observer placed a P next to any steps that required prompting for correct completion. Percent integrity was estimated by computing the number of check marks over the total number of possible intervention steps and multiplying by 100%. Average procedural integrity was 96% (range: 80-100%). The most frequent error in implementation was failing to ensure that students corrected their errors following scoring of their worksheets during independent timed practice. Teachers were provided with performance feedback following the observation to facilitate future implementation accuracy.

Criteria were empirically derived using data from 4 weeks of the single-skill probes tracking student growth in response to a standard intervention. The probes began on Friday of the same week as the March 2004 mixed skill probes and occurred every subsequent Friday for the next 3 weeks. Although the items varied between probes, the skill did not (e.g., single-digit by double-digit multiplication).

Data from the single-skill probes were recorded as digits correct/minute (dc/min) and were graphed for each individual student. A regression line was then fitted to each graph based upon ordinary least squares regression using scores and number of weeks. Therefore, the resulting slopes indicated an average increase in dc/min for each week for each child. Mean slopes for the weekly probes taken over 4 weeks during intervention were 1.36 (SD = 1.17) for second-grade, 2.63 (SD = 1.60) for third-grade, 1.60 (SD = 0.93) for fourth-grade, and 0.91 (SD = 1.35) for fifth-grade students. Standard error of slopes was computed using the formula outlined by Christ (2006), which then was used to compute the reliability of the slopes given the standard deviations. The resulting reliabilities were .98 for second grade, .99 for third grade, .97 for fourth

grade, and .98 for fifth grade. Thus, all the reliability coefficients exceeded .90 and suggested adequately reliable slopes for research.

Slope data were used to classify students as responsive or nonresponsive to the intervention by using a normative criterion. Previous research used slope data to classify children as responsive or nonresponsive to an intervention. Students were classified as responsive in previous research if they scored at or above the median slope (Vellutino et al., 1996), which has been described as an acceptable criterion (Fuchs, 2003). A similar criterion was used for the current study by identifying the slope value that represented the mean and 66th percentile for each grade. The 66th percentile was used to identify high responders because previous research examining nonresponders found that the 33rd percentile adequately differentiated student skills (Burns & Senesac, 2005). Thus, because no standards were available to compare slope rate for identifying responders, it was assumed that the percentile rank that represented the same distance from the mean at the opposite end of the distribution would adequately identify students who did respond to instruction. Mean slopes were reported in the last paragraph. The slope that represented the 66th percentile was 1.52 for second grade, 2.98 for third grade, 1.88 for fourth grade, and 1.13 for fifth grade. Therefore, students whose slope values equaled or exceeded the 66th percentile for their grades were classified as high responders.

Previous research found that children experienced greater academic outcomes when instructional material was presented at the instructional level (Burns, 2002; Daly et al., 1996, Daly et al., 1997; Gickling & Armstrong, 1978, Thompson, Gickling, & Havertape, 1983). Thus, one way to derive a range of performance that might represent the instructional level was to identify the strongest responders given intervention at a particular skill level and specify their average level of performance before intervention (i.e., at baseline). In other words, it was reasoned that students who showed the strongest growth would be students for whom the task represented an appropriate instructional match (appropriate level of difficulty). Thus, criteria for the instructional level were established using mean mathematics fluency of the high-performing group on the first of the four single-skill probes used to compute the slopes. The goal was to identify a mean fluency rate from which the highest rate of learning subsequently occurred at each grade level.

Interscorer Reliability and Procedural Integrity of Assessment Procedures

Twenty-five percent ($n = 112$) of the probes were also scored by a second person and the two sets of scores were correlated using Pearson product moment to assess interscorer reliability. The resulting coefficients for second-grade students exceeded .99 for both fluency and accuracy. Coefficients for third grade were .96 for fluency and .98 for accuracy, and fourth- and fifth-grade coefficients exceeded .99 for both fluency and accuracy. These coefficients suggest adequate interscorer reliability for the probes.

In addition, administration of assessment probes was observed by trained observers using a checklist to note the unprompted occurrence of each step required for correct probe administration (e.g., scripted instructions were read to the class, teacher checked for student understanding, students were allowed 2 min to complete problems, teacher monitored during testing by walking around the room, and papers were collected). All teachers were observed

during the initial assessment and the average percent of correctly completed steps during probe administration was 100%.

Results

Analyses were conducted so the two grade groups would be consistent with Deno and Mirkin (1977). Descriptive statistics were computed for the current samples and found mean scores for second and third grades of 19.75 (SD = 9.92) for fluency, 94.04 (SD = 8.03) for accuracy, and 627.95 (SD = 38.10) for SAT-9. Mean scores for fourth- and fifth-grade students were 32.92 dc/min (SD = 12.14) for fluency, 95.92% correct responses for 2 min (SD = 5.71) for accuracy, and 662.49 (SD = 36.43) for SAT-9 **math** standard score.

The samples were further analyzed by computing estimates of skewness and kurtosis. Skewness estimates were 4.06 for fluency, -4.26 for accuracy and -0.17 for SAT-9 for second- and third-grade students. Skewness estimates were 0.67 for fluency, -3.39 for accuracy, and -0.03 for SAT-9 for fourth- and fifth-grade students. The standard error of skewness was 0.16 for second- and third-grade students and 0.17 for fourth- and fifth-grade students. Thus, a significant positive skew was found for fluency among second- and third-grade students, and a significant negative skew was noted for accuracy for both grade groups. Kurtosis estimates for students in second and third grades were 35.52 for fluency, 32.38 for accuracy, and -0.06 for SAT-9. Kurtosis estimates were 0.23 for fluency, 17.55 for accuracy, and 0.003 for SAT-9 for students in fourth and fifth grades. Standard errors of kurtosis were 0.32 for second- and third-grade students, and 0.34 for fourth- and fifth-grade students. Thus, SAT-9 data were normally distributed, but the fluency for second- and third-graders and accuracy for both grades had leptokurtic distributions.

Given the skewness and leptokurtic distributions, a multivariate analysis of outliers was conducted using Mahalanobis D, in which significant outliers on the combined variables were identified. Five variables were entered (fluency for both probes, accuracy for both probes, and SAT-9 score), which resulted in a significant value of 20.52. Therefore, two students in Grades 2 and 3, and four students in Grades 4 and 5 were identified as outliers and were excluded from the study. Moreover, 22 students in each grade group were missing data for at least one of the five scores and were excluded from the analyses. Thus, there were 208 students in the second and third grades and 176 students in the fourth and fifth grades with acceptable data used for the analyses. The estimates of skewness and kurtosis were examined a second time after outlying data were removed. The results suggested that the SAT-9 data remained normally distributed for both groups. The fluency data demonstrated acceptable kurtosis (second and third grades = 0.20, SE = 0.34; fourth and fifth grades = 0.16, SE = 0.36), but the data remained skewed, although to a smaller degree (second and third grades = 0.61, SE = 0.17; fourth and fifth grades = 0.65, SE = 0.18). Accuracy data remained both skewed (second and third grades = -1.31, SE = 0.17; fourth and fifth grades = -1.84, SE = 0.18) and leptokurtic (second and third grades = 1.58, SE = 0.34; fourth and fifth grades = 4.57, SE = 0.36). Thus, results of analyses with the accuracy data, and fluency data to a lesser degree, should be interpreted cautiously.

Table 1 lists the descriptive statistics of the data after the outliers were removed. Two separate multivariate analyses of variance were conducted, each with two levels and three dependent variables. The levels were the grades within the grade groupings (i.e., second and third grades,

and fourth and fifth grades), and the dependent variables were the fluency, accuracy, and SAT-9 scores. A significant main effect was obtained for students in the second and third grades [$F(3, 204) = 74.65, p < .01$] and for those in fourth and fifth grades [$F(3, 172) = 15.99, p < .01$], but the average effect size was small ($d = -0.07$) for second- and third-grade students and moderate ($d = 0.51$) for fourth- and fifth-grade students. Among second- and third-grade students, grade significantly differentiated fluency [$F(1, 206) = 43.01, p < .01$], accuracy [$F(1, 206) = 9.63, p < .01$], and SAT-9 scores [$F(1, 206) = 69.54, p < .01$]. Among the fourth- and fifth-grade students, grade did not differentiate accuracy [$F(1, 174) = 0.31, p = .58$], but did differentiate fluency [$F(1, 174) = 21.16, p < .01$] and SAT-9 [$F(1, 174) = 37.68, p < .01$].

Reliability and Validity of Fluency and Accuracy Scores

Delayed alternate-form reliability estimates were computed for the fluency and accuracy measures with the Pearson product moment correlation and these results are displayed in Table 1. Fluency exceeded 0.60 for both grade groups and was considerably higher than the correlations between accuracy probes. The coefficients for the total sample were compared using Fisher's z transformation, which found that fluency scores were significantly more reliable ($z = 7.50, p < .001$) than accuracy scores.

Kendall's tau correlations of categorical data (frustration, instructional, and mastery) are listed in Table 2. Categories determined by fluency data were consistently more reliable than those derived from accuracy data, and the fourth- and fifth-grade group had higher coefficients than did the second- and third-grade students. Moreover, a percent agreement between categories was computed by determining the number of students for whom the category was the same for both assessments. Fluency categories resulted in a percent agreement of 62.4% for second- and third-grade students and 79.4% for fourth- and fifth-grade students. Accuracy categories resulted in 87.2% (second- and third-graders) and 96.8% (fourth- and fifth-graders) agreement. The tau coefficients for the total sample were again compared using Fisher's z transformation, and suggested that the categories determined by fluency scores were significantly more reliable ($z = 3.40, p < .001$) than those determined with accuracy.

Mathematics fluency and accuracy scores were correlated with SAT-9 mathematics scores with the Pearson product moment. The results of these analyses are displayed in Table 3 and suggested that scores from fourth- and fifth-grade students exceeded those of second- and third-grade students. Continuous data from the first probe were correlated with SAT-9 standard scores using a Pearson product moment correlation, and categorical data from probe one (frustration, instructional, mastery) were correlated with the continuous data of the SAT-9 standard scores using Spearman's rho. The fluency and accuracy approaches to estimating the relationship between the probe and SAT-9 scores were examined by again comparing the aggregate coefficients using Fisher's z transformation. Results indicated that the fluency scores correlated significantly higher ($z = 3.20, p < .001$) with SAT-9 scores than did accuracy scores, but categories developed from the fluency metric correlated with SAT-9 scores equally well as categories derived from accuracy scores ($z = 0.90, p > .05$).

Empirically Derived Criteria

Table 4 lists the empirically derived instructional levels. These criteria were obtained by computing the mean and 1 SD range of the starting fluency score, from the initial mathematics single-skill probes, for children whose growth slopes met or exceeded the 66th percentile. Seventy-two (35.5%) students from the second and third grades were classified as high responders, with a mean starting fluency rate of 22.76 dc/min (SD = 8.38), and 61 (34.9%) students from the fourth and fifth grades were high responders, with a mean starting fluency rate of 36.70 dc/min (SD = 12.21). An instructional level criterion was suggested for each grade grouping by computing the starting fluency score from the first of the four single-skill probes that fell 1 SD above and below the mean for each group. The resulting numbers were rounded to the nearest whole digit. Fluency scores that fell below the low end of the range would be within the frustration level and those that exceeded the highest end of the range would be within the mastery level.

Next, data from the two mixed probes from March 2004 were recategorized as frustration, instructional, or mastery using the new empirically derived criteria (see Table 4). The frequency of classifications is listed in Table 4 and was compared to category frequency of the same data with the Deno and Mirkin (1977) criteria using a Wilcoxon signed rank test. Results suggested significant changes in classifications according to criteria used for both the second and third grades [$Z(n = 208) = 10.54, p < .001$] and the fourth and fifth grades [$Z(n = 176) = 6.93, p < .001$] students. The ranking (frustration, instructional, and mastery) remained the same for 96 students in the first group and 128 in the second, but the ranking decreased (from mastery to instructional or instructional to frustration) for 112 and 48 students respectively.

Delayed alternate-form reliability was estimated for the new categorical criteria by correlating (Kendall's tau) the categories from the two probes, and these categorical data were also correlated (Spearman's rho) with the continuous data from the SAT-9 mathematics standard scores to estimate criterion-related validity. The reliability coefficients were .35 and .63 for the two grade groups, and .51 for the total sample. Percent agreement was again computed and resulted in 69.0% agreement for second- and third-graders and 92.4% for fourth- and fifth-graders. Validity coefficients were .08 and .49 for the grade groups, and .27 for the total sample. Thus, students in the fourth and fifth grades again had higher coefficients than those in second and third grades.

Finally, the mixed probe was used to classify children as frustration, instructional, or mastery using the new criteria and the average slopes from the four single-skill probes were computed for children in each category to see if faster growth rates were observed for children classified as falling in the instructional range relative to the growth of those in the frustration and mastery categories using the new criteria. In other words, the validity of the newly derived criteria was evaluated by applying the criteria to a different set of data, then comparing the average slopes of students in each category. A larger mean slope for students classified within the instructional level would support the validity of the categories. Among second- and third-graders, the mean slopes were 1.77 dc/min growth per week (SD = 1.51) for frustration, 2.01 dc/min growth per week (SD = 1.55) for instructional, and 1.55 dc/min growth per week (SD = 1.46) for mastery. Mean slopes for fourth- and fifth-grade students were 1.16 dc/min growth per week (SD = 0.99) for frustration, 1.44 dc/min growth per week (SD = 1.09) for instructional, and 1.25 dc/min growth per week (SD = 1.15) for mastery. One-way analyses of variance for these two sets of

data did not reveal significant effects [$F(2, 200) = 0.87, p = .42$, and $F(2,172) = 1.08, p = .34$, respectively]. Differences between mean scores were further analyzed by computing Cohen's (1988) d , which suggested a small average effect of 0.22 between instructional and frustration (0.16 for second- and third-graders, and 0.27 for fourth- and fifth-graders) and of 0.24 between instructional and mastery (0.31 for second- and third-graders, and 0.17 for fourth- and fifth-graders) categories. However, the average size of the effect for frustration and mastery of 0.11 was small to negligible (0.14 for second- and third-graders, and 0.08 for fourth- and fifth-graders). The mean effect sizes for slopes categorized according to the Deno and Mirkin (1977) criteria were small, but fairly similar to the new criteria when comparing frustration and instructional levels (average $d = 0.29$) and instructional and mastery (average $d = 0.16$). However, the effect between mastery and frustration levels was small, but almost twice as large (average $d = 0.20$).

Discussion

The fluency with which students solve computational problems in **math** has been recognized as an important goal of mathematics instruction and appropriate for assessment. Yet few studies have examined the consistency with which fluency scores on computational tasks are stable over time, are sensitive to intervention, and can be used to guide instructional programming. Critically, the ranges of performance provided by Deno and Mirkin (1977), an important guide for instructional programming, have not been empirically validated. The current study compared the utility and stability of fluency and accuracy scores alone for determining instructional level for mathematics. Data from students in fourth and fifth grades generally demonstrated higher coefficients than those of second and third grade students, and the aggregated fluency data were more reliable than accuracy data. Higher reliability coefficients are needed when decisions are made regarding individual students, but a coefficient of .80 from a 2-week test-retest interval suggests sufficient reliability for screening or instructional planning (Salvia & Ysseldyke, 2004). The current study used a delayed alternate-form approach, which takes multiple sources of test error into account and suggests more confidence can be asserted when interpreting the data. The coefficient for fourth- and fifth-grade students exceeded .80 for measures of mathematics fluency. The fluency coefficient for second- and third-grade students exceeded .60, but given that this was derived from a delayed alternate-form method, this could suggest some potentially useful data for instructional planning for groups of students and cautious use of data for individual students. Only categorical data based on fluency (Deno & Mirkin, 1977) appeared sufficiently reliable for screening and instructional decisions for individual students and to allow decisions about groups of students for fourth- and fifth-graders. Moreover, categorical data were more reliable for the total sample when the categories were established with fluency than with accuracy data.

Interpreting criterion-related validity coefficients with rigid categories should be avoided, but coefficients from most studies fall between .30 and .40 and coefficients larger than .60 are rare (Kaplan & Succazzo, 2001). Fluency scores were more strongly correlated with SAT-9 scores than were accuracy scores. Moreover, the coefficients from fluency scores among fourth- and fifth-grade students, and the total sample, were generally consistent with previous research (Foegen & Deno, 2001; Skiba et al., 1986), reporting coefficients between .29 and .61. The coefficients for the second- and third-grade students suggested a small to nonexistent

relationship, which is also consistent with previous research showing lower coefficients for younger children relative to older children (Foegen & Deno, 2001).

Results of the current analyses suggested that these fluency data were psychometrically preferable to the accuracy data and continuous data psychometrically outperformed the categorical data. It makes some sense that continuous data such as digits correct per minute or percent correct would lead to higher correlation coefficients than the categorical data because the scores decreased to a range of one to three categories. Thus, somewhat lower correlations would be expected. However, previous research examining the 2-week test-retest reliability of categorical data for reading, also using three categories correlated with Kenall's tau, resulted in coefficients that ranged from .82 to .89 among second-, third-, and fourth-grade students (Burns, Tucker, Frame, Foley, & Hauser, 2000). Therefore, lower reliabilities probably cannot be completely explained by the decreased range and perhaps were related to actual differences in consistency.

Implications for Practice and Research

The two instructional levels derived from the current data were somewhat similar to Deno and Mirkin's (1977), but the current standard for instructional level performance was higher relative to Deno and Mirkin's standard (17 dc/min as compared to 9 dc/min for second- and third-grade students, and 25 dc/min as compared to 19 dc/min for fourth- and fifth-grade students). Similarly, the threshold for the mastery range was higher relative to the threshold described by Deno and Mirkin (1977). Hence, the criteria for instructional level performance reported in this article were higher than those suggested by Deno and Mirkin (1977), which had a significant effect on placements of instructional categories of students. Moreover, the two approaches to categorizing student performance demonstrated comparable psychometric properties, with some notable differences in the effect sizes when comparing mean slopes of students in the mastery and frustration level. Thus, the primary advantage of the criteria found in the current study is their empirical origin. Finding an instructional level for mathematics is important because interventions could occur with children who experience mathematics difficulties until their skills reach an instructional level, at which point the child could participate in general instruction and be expected to experience improved learning outcomes (Gickling et al., 1989; VanDerHeyden & Burns, 2005). In other words, instructional level criteria can be used for instructional placement decisions (Shapiro, 2004), and the criteria used here demonstrated a significant effect on the placement decision, which warrants additional research.

Limitations and Directions for Future Research

Several limitations may affect the utility of these findings. These data were collected within one school district, and the degree to which these findings would apply to students in other districts is unknown. Moreover, the fluency data were minimally skewed after removing outlier scores, and the accuracy data were both skewed and leptokurtic. Thus, risk of Type I error might have been inflated. Further, data were analyzed according to two grade groupings to assure consistency with Deno and Mirkin (1977), but small to moderate effects were noted between grades within the same group. Future research may thus wish to examine grade-specific criteria instead of grade groups. Accuracy was one of the variables considered in this study, and yet, technically,

the way that accuracy was measured also included a fluency component. That is, all probes were administered using standardized instructions that included a 2-min timed administration period. Hence, our accuracy estimate was actually percentage correct responses in 2 min. The degree to which the same findings would have been obtained had a more pure measure of accuracy been used is unknown. Moreover, the study deviated slightly from standardized procedures, the effect of which is unknown. Finally, the stability of performance on the mixed-skill probes over time was likely negatively affected by the fact that children were participating in daily intervention on related skills during the same time period. To the degree that children's scores were affected by intervention on related skills, the reliability estimates reported in this article may be underestimated.

Math presents, perhaps, a unique challenge in curriculum-based assessment and measurement. With **math**, the task can vary substantially not only in terms of difficulty, but along many other dimensions as well. Therefore, when measuring **math** performance, it is important to specify the stimulus conditions under which the data were obtained. In this study, slope data were computed using single-skill probes that differed in content between grades and classes within grades. This approach was used because our purpose was to attempt to identify growth given controlled intervention within a particular grade-level appropriate skill and then to determine what level of fluency at baseline was associated with strongest growth during subsequent intervention. The approach was guided by pragmatism and the desire to identify a criterion that had functional utility in determining what difficulty level would constitute an instructional match. Thus, potential differences in findings between skills were not examined and require further empirical investigation.

Data presented in the current study suggested that perhaps the main advantage of the current criteria over previously suggested fluency criteria is the empirical, as opposed to intuitive, nature with which they were developed. Yet, empirically deriving these instructional level criteria is only the first step of research regarding their use in mathematical assessment. Additional research is needed to examine whether behavioral outcomes that have been associated with the instructional level for reading (e.g., increased time on task and task completion, enhanced comprehension) are also found with these mathematics categories. Moreover, research is needed to explore the degree to which various stimulus sets relative to curricular expectations and student ability produce meaningful decision making for screening, progress monitoring, and even placement purposes. Given the importance of an instructional match between student skill and curriculum (Ysseldyke & Christenson, 2002), and that assessment of the instructional level and subsequent interventions based on those data have been consistently linked to improved student outcomes in reading (Burns, 2002; Burns, Dean, & Foley, 2004; Thompson et al., 1983), additional research in mathematics appears warranted.

Supplementary Material

For a further discussion of implications for practice, go to www.nasponline.org/publications/sprsupplemental.html.

References

Algozzine, R., Ysseldyke, J., & Elliott, J. (1997). *Strategies and tactics for effective instruction* (2nd ed.). Longmont, CO: Sopris West.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington DC: American Psychological Association.

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Upper Saddle River, NJ: Prentice Hall.

Badian, N. (1999). Persistent arithmetic, reading or arithmetic, or reading disability. *Annals of Dyslexia*, 49, 45-70.

Betts, E. A. (1946). *Foundations of reading instruction*. New York: American Book.

Binder, C. (1996). Behavioral fluency: Evolution of a new paradigm. *Behavior Analyst*, 19, 163-197.

Burns, M. K. (2002). Utilizing a comprehensive system of assessment to intervention using curriculum-based assessments. *Intervention in School and Clinic*, 38, 8-13.

Burns, M. K. (2004). Empirical analysis of drill ratio research: Refining the instructional level for drill tasks. *Remedial and Special Education*, 25, 167-175.

Burns, M. K., Dean, V. J., & Foley, S. (2004). Preteaching unknown key words with incremental rehearsal to improve reading fluency and comprehension with children identified as reading disabled. *Journal of School Psychology*, 42, 303-314.

Burns, M. K., & Senesac, B. K. (2005). Comparison of dual discrepancy criteria for diagnosis of unresponsiveness to intervention. *Journal of School Psychology*, 43, 393-406.

Burns, M. K., Tucker, J. A., Frame, J., Foley, S., & Hauser, A. (2000). Interscorer, alternate-form, internal consistency, and test-retest reliability of Gickling's model of curriculum-based assessment for reading. *Journal of Psychoeducational Assessment*, 18, 353-360.

Christ, T. J. (2006). Short term estimates of growth using curriculum-based measurement of oral reading fluency: Estimating standard error of the slope to construct confidence intervals. *School Psychology Review*, 35, 128-133.

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

Daly, E. J., III, & Martens, B. K. (1994). A comparison of three interventions for increasing oral reading performance: Application of the instructional hierarchy. *Journal of Applied Behavior Analysis*, 27, 459-469.

Daly, E. J., III, Martens, B. K., Kilmer, A., & Massie, D. (1996). The effects of instructional

match and content overlap on generalized reading performance. *Journal of Applied Behavioral Analysis*, 29, 507-518.

Daly, E. J., III, & McCurdy, M. (2002). Getting it right so they can get it right: An overview of the special series. *School Psychology Review*, 31, 453-458.

Daly, E. J., III, Witt, J. C., Martens, B. K., & Dool, E. J. (1997). A model for conducting a functional analysis of academic performance problems. *School Psychology Review*, 26, 554-574.

Deno, S. L., & Mirkin, P. K. (1977). *Data-based program modification: A manual*. Reston, VA: Council for Exception Children.

Enggren, P., & Kovaleski, J. F. (1996). *Instructional assessment*. Harrisburg: Instructional Support System of Pennsylvania.

Espin, C., Deno, S. L., Maruyama, G., & Cohen, C. (1989). *The basic academic skills samples (BASS): An instrument for the screening and identification of children at risk for failure in regular education classrooms*. Minneapolis: University of Minnesota, Special Education Program.

Foegen, A., & Deno, S. L. (2001). Identifying growth indicators for low-achieving students in middle school mathematics. *Journal of Special Education*, 35, 4-16.

Fuchs, D., Fuchs, L., Mathes, P. G., & Simmons, D. C. (1997). Peer-assisted learning strategies: Making classrooms more responsive to diversity. *American Educational Research Journal*, 34, 174-206.

Fuchs, L. S. (2003). Assessing intervention responsiveness: Conceptual and technical issues. *Learning Disabilities: Research & Practice*, 18, 172-186.

Fuchs, L. S., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research and Practice*, 13, 204-219.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989). Computers and curriculum-based measurement: Effects of teacher feedback systems. *School Psychology Review*, 18, 112-125.

Fuchs, L. S., Fuchs, D., Phillips, N. B., Hamlett, C. L., & Karns, K. (1995). Acquisition and transfer effects of classwide peer-assisted learning strategies in mathematics for students with varying learning histories. *School Psychology Review*, 24, 604-620.

Fuchs, L. S., Fuchs, D., Prentice, K., Burch, M., Hamlett, C. L., Owen, R., Hosp, M., & Jancek, D. (2003). Explicitly teaching for transfer: Effects on third-grade students' mathematical problem solving. *Journal of Educational Psychology*, 95, 293-305.

Gickling, E. E., & Armstrong, D. L. (1978). Levels of instructional difficulty as related to on-

task behavior, task completion, and comprehension. *Journal of Learning Disabilities*, 11, 559-566.

Gickling, E., & Rosenfield, S. (1995). Best practices in curriculum-based assessment. In A. Thomas, & J. Grimes (Eds.), *Best practices in school psychology III* (pp. 587-595). Washington, DC: National Association of School Psychologists.

Gickling, E. E., Shane, R. L., & Croskery, K. M. (1989). Developing math skills in low-achieving high school students through curriculum-based assessment. *School Psychology Review*, 18, 344-356.

Gickling, E., & Thompson, V. (1985). A personal view of curriculum-based assessment. *Exceptional Children*, 52, 205-218.

Gravois, T. A., & Gickling, E. E. (2002). Best practices in curriculum-based assessment. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (Vol. 2, pp. 885-898). Bethesda, MD: National Association of School Psychologists.

Greenwood, C. R. (1991). Classwide peer tutoring: Longitudinal effects on the reading, language, and mathematics achievement of at-risk students. *Journal of Reading, Writing, & Learning Disabilities International*, 7, 105-123.

Haladyna, T. M. (1998). Review of the Stanford Achievement Test (9th ed.). In J. C. Impara & B. S. Plake (Eds.), *The thirteenth mental measurements yearbook* (pp. 928-930). Lincoln: The University of Nebraska--Lincoln.

Harcourt Brace Educational Measurement. (1996). *Stanford Achievement Test, Ninth Edition*. San Antonio, TX: Author.

Harcourt Publishing. (2003). *Harcourt Math*. Orlando, FL: Author.

Hintze, J. M., Christ, T. J., & Keller, L. A. (2002). The generalizability of CBM survey-level mathematics assessments: Just how many samples do we need? *School Psychology Review*, 31, 514-528.

Kaplan, R. M., & Saccuzzo, D. P. (2001). *Psychological testing: Principles, applications, and issues* (5th ed.). Belmont, CA: Wadsworth/Thomson Learning.

Manzo, K. K., & Galley, M. (2003). Math climbs, reading flat on '03 NAEP. *Education Week*, 23(12), 1-18.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.

Murphy, K. R., & Davidshofer, C. O. (2001). *Psychological testing principles and applications*

(5th ed.). Upper Saddle River, NJ: Prentice Hall.

National Council of Teachers of Mathematics. (2000). Principles and standards for school mathematics. Reston, VA: Author.

Reschly, D. J. (1996). Functional assessments and special education decision making. In W. Stainback & S. Stainback (Eds.), *Controversial issues confronting special education: Divergent perspectives* (2nd ed., pp. 115-128). Boston: Allyn & Bacon.

Rivera, D. M., & Bryant, B. R. (1992). Mathematics instruction for students with special needs. *Intervention in School & Clinic*, 28, 71-86.

Salvia, J., & Ysseldyke, J. E. (2004). *Assessment* (9th ed.). Boston: Houghton Mifflin.

Saxon Publishers. (2004). *Saxon Math*. Norman, OK: Author.

Shapiro, E. S. (1992). Use of Gickling's model of curriculum-based assessment to improve reading in elementary age students. *School Psychology Review*, 21, 168-176.

Shapiro, E. S. (2004). *Academic skill problems: Direct assessment and intervention* (3rd ed.). New York: Guilford Press.

Shapiro, E. S., & Ager, C. (1992). Assessment of special education students in regular education programs: Linking assessment to instruction. *Elementary School Journal*, 92, 283-296.

Shinn, M. R. (Ed.). (1989). *Curriculum-based measurement: Assessing special children*. New York: Guilford Press.

Shinn, M. R., Collins, V. L., & Gallagher, S. (1998) Curriculum-based measurement and its use in a problem solving model with students from minority backgrounds. In S. N. Elliot & J. C. Witt (Series Eds.) & M. R. Shinn (Vol. Ed.), *Advanced applications of curriculum-based measurement*. New York: Guilford Press.

Skiba, R., Magnusson, D., Marston, D., & Erickson, K. (1986). *The assessment of mathematics performance in special education: Achievement tests, proficiency tests, or formative evaluation*. Minneapolis: Minneapolis Public Schools.

Thompson, V. P., Gickling, E., & Havertape, J. F. (1983). The effects of medication and curriculum management on task-related behaviors of attention deficit disordered and low achieving peers. *Monographs in behavioral disorders: Severe behavior disorders of children and youth* (Series No. 6). Council for Children with Behavioral Disorders, Arizona State University.

Thurber, R. S., Shinn, M. R., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based measures. *School Psychology Review*, 31, 498-513.

Tindal, G., Germann, G., & Deno, S. L. (1983). *Descriptive research on the Pine County norms*;

A compilation of findings (Res. Rep. No. 132). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Tucker, J. A. (1985). Curriculum-based assessment: An introduction. *Exceptional Children*, 52, 199.-204.

VanDerHeyden, A. M., & Burns, M. K. (2005). Using curriculum-based assessment and curriculum-based measurement to guide elementary mathematics instruction: Effect on individual and group accountability scores. *Assessment for Effective Intervention*, 30(3), 15-29.

VanDerHeyden, A. M., Witt, J. C., & Naquin, G. (2003). Development and validation of a process for screening referrals to special education. *School Psychology Review*, 32, 204-227.

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S. G., Chen, R., Pratt, A., & Denkla, M. B. (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.

Witt, J. C., Daly, E., & Noell, G. (2000). *Functional assessments*. Sopris West: Longmont, CO.

Ysseldyke, J. Dawson, P. Lehr, C., Reschly, D., Reynolds, M., & Telzrow, C. (1997). *School psychology: A blueprint for training and practice*. Bethesda, MD: National Association of School Psychologists.

Ysseldyke, J. E., & Christenson, S. (2002). *Functional assessment of academic behavior*. Longmont, CO: Sopris West.

Date Received: 11/21/05

Date Accepted: 1/26/06

Action Editor: John Hintze

Matthew K. Bums, PhD, is an Associate Professor of Educational Psychology at the University of Minnesota. His current research interests include curriculum-based assessment, interventions for academic difficulties, response to intervention, and problem-solving teams.

Amanda M. VanDerHeyden, PhD, is an Assistant Professor at University of California at Santa Barbara. She has authored more than 25 articles and chapters and has worked as a national trainer to implement data-driven practices in schools. She is Associate Editor for *Journal of Behavioral Education* and serves on the editorial boards for *Journal of Early Intervention* and *School Psychology Review*.

Cynthia L. Jiban is a doctoral candidate in educational psychology (special education) at the University of Minnesota. Her current research interests include progress monitoring, the role of reading in mathematics achievement, and preparing teachers of students who struggle

academically.

APPENDIX SKILL SEQUENCE 2003-2004

Second-Grade Skill

1. Addition facts 0-20
2. Subtraction facts 0-9
3. Subtraction facts 0-12
4. Subtraction facts 0-15
5. Subtraction facts 0-20
6. Mixed subtraction or addition 0-20
7. Fact families addition and subtraction 0-20
8. Two-digit addition without regrouping
9. Two-digit addition with regrouping
10. Two-digit subtraction without regrouping
11. Two-digit subtraction with regrouping
12. Three-digit addition without and with regrouping
13. Three-digit subtraction without and with regrouping
14. Second-grade monthly **math** probe

Third-Grade Skill

1. Addition and subtraction facts 0-20
2. Fact families addition and subtraction 0-20
3. Three-digit addition without and with regrouping
4. Three-digit subtraction without and with regrouping
5. Two- and three-digit addition and subtraction with and without regrouping

6. Multiplication facts 0-9
7. Division facts 0-9
8. Fact families multiplication and division 0-9
9. Add or subtract fractions with like denominators
10. Single digit multiplied by double or triple digit without regrouping
11. Single digit multiplied by double or triple digit with regrouping
12. Single digit divided into double or triple digit without remainders
13. Add and subtract decimals to the hundredths

Fourth-Grade Skill

1. Multiplication facts 0-12
2. Division facts 0-12
3. Fact families multiplication or division 0-12
4. Single digit multiplied by double digit with and without regrouping
5. Double digit multiplied by double digit without regrouping
6. Double digit multiplied by double digit with regrouping
7. Single-digit divisor into double-digit dividend without remainders
8. Single-digit divisor into double-digit dividend with remainders
9. Single- and double-digit divisor into single- and double-digit dividend with remainders
10. Add or subtract fractions with like denominators no regrouping
11. Multiply multidigit numbers by two numbers
12. Add and subtract decimals to the hundredths

Fifth-Grade Skill

1. Multiplication facts 0-12

2. Division facts 0-12
3. Fact families multiplication or division 0-12
4. Multiply two- and three-digit with and without regrouping
5. Single-digit divisor divided into double-digit dividend with remainders
6. Single-digit divisor divided into double- and triple-digit dividend with remainders
7. Reduce fractions to simplest form
8. Add or subtract proper fractions or mixed numbers with like denominators with regrouping
9. Add or subtract decimals
10. Multiply or divide decimals
11. Double-digit divisor into four-digit dividend
12. Multiply and divide proper and improper fractions

Matthew K. Burns

University of Minnesota

Amanda M. VanDerHeyden

University of California at Santa Barbara

Cynthia L. Jiban

University of Minnesota

The data presented in this article were collected when the second author was affiliated with the Vail School District in Vail, Arizona.

Correspondence regarding this article should be addressed to Matthew Burns, University of Minnesota, 346 Elliott Hall, 75 E. River Road, Minneapolis, MN 55455; e-mail: burns258@umn.edu

Table 1 Means, Standard Deviations, and Correlation Coefficients for Fluency and Accuracy Scores

Fluency		Accuracy			
Probe 1	Probe 2	Probe 1	Probe 2		
Group	M	SD	M	SD	r
	M	SD	M	SD	r

Second grade 22.7 7.9 22.1 99.8 .71* 95.8 5.6 96.1 5.4 .49*
 Third grade 16.4 5.4 18.7 6.8 .42* 93.3 6.0 93.1 7.1 .19
 Second and third grades 19.8 7.5 20.5 8.7 .64* 94.7 5.9 94.7 6.4 .36*
 third grades
 Fourth grade 30.1 10.0 29.2 11.9 .78* 96.6 4.1 96.7 4.4 .48*
 Fifth grade 38.1 12.9 38.2 14.2 .88* 96.3 4.2 97.5 3.4 .57*
 Fourth and fifth grades 33.5 12.0 32.9 13.6 .85* 96.5 4.1 97.0 4.0 .50*
 fifth grades
 Total sample 26.0 11.9 26.2 12.8 .84* 95.5 5.2 95.8 5.6 .42*

*p < .01.

Table 2 Number and Percentage of Scores Categorized as Frustration, Instructional, and Mastery and Correlation Coefficients

Probe 1 Probe 2
 Frust. Inst. Mast. Frust. Inst.
 Group N % N % N % N % N %

 Second and third grades
 Fluency 13 6.3 98 47.1 97 46.6 18 8.7 89 42.8
 Accuracy 0 0.0 17 8.2 191 91.8 2 1.0 17 8.2
 Fourth and fifth grades
 Fluency 19 10.8 108 61.4 49 27.8 32 18.2 90 51.1
 Accuracy 0 0.0 4 2.3 172 97.7 0 0.0 04 2.3
 Total sample
 Fluency 32 8.3 206 53.6 146 38.0 50 13.0 179 46.6
 Accuracy 0 0.0 21 05.5 363 94.5 2 0.5 21 5.5

Probe 2
 Mast.
 Group N % [tau]

Second and third grades
 Fluency 101 48.6 .42*
 Accuracy 189 90.9 .27*
 Fourth and fifth grades
 Fluency 54 30.7 .71*
 Accuracy 172 97.7 .49*
 Total sample
 Fluency 155 40.4 .60*
 Accuracy 361 94.0 .33*

Note. Frust. = frustrational level; Inst. = instructional level; Mast. = mastery level.

*p < .01.

Table 3 Criterion-Related Validity Coefficients Between Mathematics Probes and SAT-9 Mathematics Test

SAT-9 Standard
 Score
 Group M SD [r.sup.1] [r.sup.2]

 Second and third grades 627.5 37.4 .16 .21*
 Fourth and fifth grades 664.5 36.0 .60* .27*

Total 644.5 41.1 .55* .28*

Group $[[\rho].sup.1]$ $[[\rho].sup.2]$

Second and third grades .06 .20*

Fourth and fifth grades .52* .18

Total .14* .23*

Note. $[r.sup.1]$ = Pearson r between SAT-9 score and mathematics fluency score; $[r.sup.2]$ = Pearson r between SAT-9 score and mathematics accuracy score; $[p.sup.1]$ = Spearman rho between SAT-9 score and fluency category score; $[p.sup.2]$ = Spearman rho between SAT-9 score and accuracy category score.

*p < .01.

Table 4 Empirically Derived Fluency Criteria With Accompanying Reliability, Validity, and Frequency

Digits Correct

Data per Minute for

Frust. Inst. Mast. Instructional

Group N % N % N % Level

Second and third grades 46 22.1 145 69.7 17 8.2 14-31

Fourth and fifth grades 38 21.6 118 67.0 20 11.4 24-49

Reliability Validity

Group ($[\tau]$) ($[\rho]$)

Second and third grades .35* .08

Fourth and fifth grades .63* .50*

Note. Frust. = frustrational level; Inst. = instructional level; Mast. = mastery level.

*p < .01.

Questia Media America, Inc. www.questia.com

Publication Information: Article Title: Assessing the Instructional Level for Mathematics: A Comparison of Methods. Contributors: Matthew K. Burns - author, Amanda M. Vanderheyden - author, Cynthia L. Jiban - author. Journal Title: School Psychology Review. Volume: 35. Issue: 3. Publication Year: 2006. Page Number: 401+. COPYRIGHT 2006 National Association of School Psychologists; COPYRIGHT 2006 Gale Group