Generalizability Theory Analysis of CBM Maze Reliability in Third- Through Fifth-Grade Students

Sterett H. Mercer
University of British Columbia
Brad A. Dufrene
Kimberly Zoder-Martell
Lauren Lestremau Harpole
Rachel R. Mitchell
John T. Blaze
The University of Southern Mississippi

Author Note

Sterett H. Mercer, Department of Educational & Counselling Psychology and Special Education, University of British Columbia; Brad A. Dufrene, Kimberly Zoder-Martell, Lauren Lestremau Harpole, Rachel R. Mitchell, and John T. Blaze, Department of Psychology, The University of Southern Mississippi.
We would like to thank Max Woodliff, Chelsi Clark, Aimee Maldonado, Abby Lambert, and Leila Mullooly for assistance with data collection and assessment scoring.
Corresponding author: Sterett H. Mercer, University of British Columbia, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada; email: sterett.mercer@ubc.ca

Abstract

Despite growing use of CBM maze in universal screening and research, little information is available regarding the number of CBM maze probes needed for reliable decisions. The current study extends existing research on the technical adequacy of CBM maze by investigating the number of probes and assessment durations (1-3 minutes) needed for reliable relative (e.g., rank-ordering students) and absolute (e.g., comparing a specific score to a cutoff) decisions. Nine CBM maze probes were administered to 272 students in third through fifth grades. Results suggested that the number of probes needed for reliable relative and absolute decisions varied by grade, with assessments in fifth grade exhibiting the highest reliability (at least two probes needed for both types of decisions). In addition, declining gains in reliability appeared to occur as assessment duration increased. Implications of the findings for universal screening and future research are discussed.

Keywords: curriculum-based measurement, generalizability theory, reading, maze fluency

Generalizability Theory Analysis of CBM Maze Reliability in Third- Through Fifth-Grade Students

Although assessment of oral reading fluency (ORF) is the primary curriculum-based measure (CBM) of reading used in research and practice (Reschly, Busch, Betts, Deno, & Long, 2009), CBM maze is growing in popularity as an additional measure. On typical CBM maze tasks, students are presented with a passage of approximately 250 words in which every seventh word has been deleted and replaced with three options. The increased use of CBM maze is partly due to efficiency of administration and because teachers perceive it as more reflective of reading comprehension than ORF (Fuchs & Fuchs, 1992; Fuchs, Fuchs, & Maxwell, 1988). Several studies support the alternate-form reliability, sensitivity to growth, and predictive validity of CBM maze (e.g., Espin, Wallace, Lembke, Cambell, & Long, 2010; Graney, Martínez, Missall, & Aricak, 2010; Shin, Deno, & Espin, 2000), and CBM maze has been suggested as a time-efficient option to ORF for universal screening given the possibility of group and/or computer administration (Fuchs & Fuchs, 1992; Graney et al., 2010). Prior to more widespread adoption of CBM maze for these purposes, additional information regarding the technical adequacy of CBM maze is needed.

In research and practice, reading benchmark procedures differ in the number of CBM maze probes administered; for example, only one probe is administered per benchmark in the System to Enhance Educational Performance [STEEP] model (see http://www.isteep.com), but three probes per benchmark were administered in Deno et al. (2009). Additionally, the duration of most CBM maze probes is three minutes, yet few studies have explicitly examined the impact of probe duration on technical adequacy. Determining the optimal number of probes to be

administered and ideal probe durations is complicated because little information is available regarding how the these factors impact the reliable differentiation of absolute levels of student performance.

Research on ORF has found that reliability varies depending on the intended use of the assessment (i.e., relative vs. absolute decisions) and as a function of variability in passage difficulty (Hintze, Owen, Shapiro, & Daly, 2000; Poncy, Skinner, & Axtell, 2005), and similar concerns may apply to the reliability of CBM maze. Specifically, estimates of alternate-form reliability have been used to support the reliability of CBM maze (e.g., Espin et al., 2010; Graney et al., 2010); however, alternate-form reliability primarily indicates that individuals are rank ordered in a similar manner across the measures (Shavelson & Webb, 1991). Consequently, adequate estimates of alternate-form reliability suggest that CBM maze is appropriate for relative decisions, such as identifying the lowest 20% of students in a class as in need of intervention.

Because alternate-form reliability and other correlational measures indicate similarity in rank ordering, these estimates largely are unaffected by variations in passage difficulty (i.e., if all students tend to perform better or worse on certain probes). Recent research has indicated that ORF passages, even when they have similar readability estimates, demonstrate differences in passage difficulty that can impact the comparability of scores across different probes (Ardoin, Williams, Christ, Klubnik, & Wellborn, 2010). Assuming that CBM maze probes have similar differences in difficulty, limited information is available regarding the suitability of CBM maze for absolute decisions, such as comparing a student's score to a specific cut score or to the student's score on a prior occasion. Given that many typical uses of CBM involve absolute decisions (e.g., comparing students' CBM scores to benchmarks and evaluating students' CBM scores relative to scores on prior occasions to determine response to intervention; Christ,

Johnson-Gros, & Hintze, 2005), developing reliability-based guidelines for the number of probes to be administered and optimal probe duration is of vital importance.

Research suggests that differences in ORF passage difficulty limit reliability for absolute decisions and that field testing of ORF probes to reduce passage variability can improve this form of reliability (Christ & Ardoin, 2009; Poncy et al., 2005). In Poncy et al. (2005), absolute decisions, based on a single ORF probe from a commercially-available probe set, exhibited reliability of .81, which is considerably below the recommended reliability of .90 for important decisions related to specific test scores (Nunnally, 1978). After poorly-performing probes, identified by comparisons of average scores on specific probes relative to the overall average, were excluded from the probe set, reliability for absolute decisions based on single ORF probes increased to .89 (Poncy et al., 2005). Provided similar differences in probe difficulty are found on CBM maze, it is possible that field testing methods to reduce this variability (Christ & Ardoin, 2009; Poncy et al., 2005) can be applied to CBM maze to improve absolute decisions.

The purpose of this study is to examine the reliability of CBM maze as a function of the number of probes administered, assessment duration (1, 2, or 3 minutes), and the purpose of the assessment (i.e., relative vs. absolute decisions). By examining reliability of CBM maze in more detail, we sought to expand knowledge regarding the technical adequacy of CBM maze beyond the alternate-form reliability estimates reported in prior studies. In addition, by determining the number of probes needed for reliable relative and absolute decisions, we hoped to provide information useful for practitioners considering the use of CBM maze as an efficient adjunct or replacement for ORF in universal screening and progress monitoring.

Method

Participants

Participants (n = 272) were third- (n = 84), fourth- (n = 91), and fifth- (n = 97) grade students without current special education eligibility from two public elementary schools in the southern United States. At one school, 93% of students were identified in school records as African American and 6% Caucasian, with 66% of students qualifying for free or reduced lunch. At the other school, 72% of students were identified as Caucasian and 28% African American, with 85% qualifying for free or reduced lunch.

Measures

CBM Maze.

Nine probes at each grade level were selected for administration from the AIMSweb maze probe set (Edformation, 2003). The number of probes was selected by balancing the goal of adequately assessing variation in probe difficulty, which requires including as many different probes as possible, with feasibility of administration based on time constraints at participating schools. The probes included approximately 250 words, with every seventh word deleted and replaced with a near distractor (syntactically similar), a far distractor (syntactically and semantically different), and the correct word. Students read the passages silently and circled the correct words. Correct maze choices (CMC) during the three-minute administration were scored and are reported as CMC per minute. Alternate-form reliability estimates (e.g., .82 using AIMSweb passages with fourth- and fifth-grade students; Graney et al., 2010), sensitivity to growth across the academic year, and correlations with standardized reading and language arts assessments have been used to support the reliability and validity of CBM maze (Espin et al., 2010; Graney et al., 2010).

Procedure

Students completed nine, grade-level CBM maze probes over three consecutive days, with three probes administered per day. The order of probe administration across students was randomized. Doctoral students in school psychology with formal coursework and prior experience in CBM group-administered the probes in classrooms. The doctoral students had been previously trained to a 90% agreement criterion for scoring a variety of curriculum-based measures. Students were instructed to read silently and circle correct word choices for three minutes. Following each minute, they were instructed to make a slash mark to indicate progress

in the passage. Aside from the instruction to make slash marks at each minute, standard administration instructions and scoring procedures were used (Shinn & Shinn, 2002). CMC at one minute, two minutes, and three minutes were scored. Inter-scorer agreement was calculated on 25% of all probes administered by dividing the number of agreements by the total number of agreements plus disagreements and multiplying by 100. Average agreement was 99%.

Data Analyses

Reliability estimates were calculated based on generalizability theory (G theory; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). As compared to classical test theory (Lord & Novick, 1968) which considers observed scores to be comprised of a true score plus measurement error, G theory considers observed scores to be attributable to individual differences plus many systematic and random sources of error variance (e.g., rater, item, occasion) that can be separated. In this study, overall variance in students' scores across the nine probes was considered to be a function of individual differences in student reading skill (*person*), variability in overall probe difficulty (*probe*), and differences in performance on specific probes by individual students (*residual*). Based on data from all nine probes in each grade, these variance components (*person, probe, residual*) were estimated as random effects in linear mixed models using restricted maximum likelihood estimation (REML) with the 'lmer' function of the 'lme4' package (Bates, Maechler, & Bolker, 2011) in R (R Development Core Team, 2010). The REML estimator was used instead of estimation of variance components using analysis of variance because simulation studies have supported REML as less biased (Marcoulides, 1993). Separate analyses were conducted for each assessment duration (i.e., one, two, and three minutes) in each grade (i.e., total of 9 models). These variance components (based on results from all nine probes) were used to calculate reliability estimates, using formulas described in the

results section, to simulate what reliability would be based on averages of one to five probes. This process is similar to using the Spearman-Brown prophecy formula to determine the optimal number of items to yield acceptable reliability in classical test theory (Shavelson & Webb, 1991). Although the formulas can be used to simulate what reliability would be with more than five probes administered, we did not present results for more than five probes because administering this many probes per occasion or benchmark is unlikely to be perceived as acceptable or feasible by researchers and practitioners.

Results

Table 1 includes descriptive statistics (i.e., means and standard deviations by probe, grade level, and assessment duration). Table 2 presents estimates of each variance component based on the separate analyses for probes of different durations by grade level. To calculate reliability estimates for relative decisions (i.e., rank-ordering students), the variance component for between-student differences (*person*) was divided by the total amount of relevant variance (*person + residual*). Variability due to overall differences in performance across probes (*probe*) was not included in the calculations because it would not affect the rank ordering of students. To simulate what reliability would be with different numbers of probes administered, the error variance (*residual*) was divided by the number of probes to be used in the decision (range: one to five) and used in repeated reliability calculations. For example, to calculate reliability for relative decisions based on three, 3-minute probes in third grade, the *person* variance component (19.07) was divided by the total of the relevant variance components, i.e., $person + residual \div n_{probes}$ (19.07 + 11.77 ÷ 3 = 22.99), yielding a reliability estimate of .83 (19.07 ÷ 22.99). This information is presented in Figure 1.

In general, reliability was greater for probes of longer durations, although the increase in reliability appeared to diminish as duration was increased from 2 to 3 minutes in third and fifth grades. In addition, overall reliability was higher in fifth grade. Administration of three probes (3 min.) would yield acceptable reliability for low-stakes decisions (e.g., screening or research, $\geq$ .80; Nunnally, 1978) in third (.83) and fourth (.83) grade, and two probes (3 min.) would yield acceptable low-stakes reliability in fifth grade (.87). Administration of five probes (3 min.) would yield excellent reliability for high-stakes decisions (e.g., program eligibility, $\geq$ .90; Nunnally, 1978) in third (.90) and fourth (.90) grade, and three probes would yield excellent reliability in fifth grade (.91). Reliability for relative decisions is important to consider in activities such as identifying the lowest performing students in a class or grade.

Figure 2 includes reliability estimates for absolute decisions. Calculation of these estimates differed by including *probe* as an additional source of error variance in the denominator, and reliability for one to five probes administered was simulated by dividing between-student variance (*person*) by the total variance ($person + probe \div n_{probes} + residual \div n_{probes}$), with *probe* and *residual* both divided by the number of probes to be administered. Depending on the duration of the probe, overall mean differences in performance across probes (*probe*) accounted for 7 to 19% of the variance in scores. Again, probes of longer duration were more reliable, with smaller gains in reliability as duration increased from 2 to 3 minutes in third and fifth grades, and greater reliability was found in fifth grade as compared to the other grades. The number of probes needed to achieve acceptable reliability to make low-stakes absolute decisions varied by grade: three in third grade (.80), four in fourth grade (.82), and two in fifth grade (.82). For excellent reliability to make high-stakes absolute decisions, more than five probes would need to be administered in third and fourth grade, and four probes would need to

be administered in fifth grade (.90). These reliability estimates are important to consider in activities such as comparing a student's score to a specific benchmark score or the student's score on another occasion.

In addition to reliability estimates, the standard error of measurement (SEM) can be helpful when considering the number of probes to be administered for specific decisions. SEM can be calculated by taking the square root of the error variance, which varies depending on the type of decision and the number of probes to be administered. Specifically, the error variance for relative decisions is $residual \div n_{probes}$ and is $probe \div n_{probes} + residual \div n_{probes}$ for absolute decisions. SEMs for administrations of one to five probes (3-minute) by decision type and grade are presented in Table 3. These SEMs can be used to create confidence intervals around benchmark or other target scores. For example, the SEM for absolute decisions based on two probes in fourth grade is approximately 3 CMC, and a 68% confidence interval can be formed around specific scores by adding and subtracting the SEM. If the target benchmark score were within $\pm$ 3 CMC of a fourth-grade student's average CMC on two probes, then it would be helpful to administer another maze probe to reduce the potential impact of measurement error on an important absolute decision, i.e., whether or not the student's performance is above the benchmark score.

Discussion

The primary purpose of this study was to broaden the knowledge base regarding the reliability of CBM maze assessments beyond the alternate-form reliability estimates reported in prior studies. Consistent with similar studies on ORF (e.g., Christ & Ardoin, 2009; Poncy et al., 2005), a sizable proportion of the variance in scores was attributable to overall mean differences in performance across probes, most likely reflecting differences in difficulty across probes. In

contrast to ORF, the importance of passage difficulty has been largely overlooked in research on CBM maze. The systematic differences in performance across probes in this study were found on commercially available probes that have undergone prior field testing (Howe & Shinn, 2002); consequently, additional development and field testing of CBM maze probes is recommended prior to more widespread use of CBM maze for absolute decisions (e.g., comparing specific scores to cutoffs or progress monitoring for individual students). Although not explored in this study, the possibility that selecting a subset of the least variable probes (e.g., Christ & Ardoin, 2009) or statistical equating (e.g., Betts, Pickard, & Heistad, 2009) could reduce the number of CBM maze probes needed for reliable decisions should be explored in future research.

In practice, administration of one to three CBM maze probes, alone or in combination with several ORF probes, is common. In research, administration of one CBM maze probe per occasion (e.g., Ardoin et al, 2004; Graney et al., 2010) or the average of two probes (Espin et al, 2010) is common. Based on the analyses in this study, however, administration of three 3-minute probes per occasion in third and fourth grade and two probes per occasion in fifth grade should be considered the minimum requirement for reliable scores for relative decisions (e.g., screening or rank-ordering students) or correlational analyses (e.g., estimates of criterion-related validity). Given that most of the studies examining the validity of CBM maze have employed only one probe per measurement occasion, it is possible that estimates of validity may be attenuated by measurement error. For this reason, validity of CBM maze, as measured by the mean of multiple probe administrations, needs to be re-examined.

The majority of studies on CBM maze have used 3-minute probes (e.g., Ardoin et al., 2004; Espin et al., 2010; Graney et al., 2010), although some studies have used longer durations for CBM maze probes. For example, Brown-Chidsey, Davis, & Maya (2003) examined the

reliability and validity of 10-minute maze probes. In the current study, assessment duration was capped at 3 minutes for individual probes. Results suggest, however, that there may be diminishing returns in reliability as administration time increases. For this reason, administration of more than one probe and additional field testing to reduce variability related to probe difficulty appear to be the primary strategies to increase reliability.

This study has several limitations that should be considered. Data were only collected on third through fifth grade students; consequently, the generalizability of findings to other grades is unknown. This limitation is important considering that some prior research has focused on the technical adequacy of CBM maze in secondary students (e.g., Espin et al., 2010). Because CBM maze is included as the primary reading assessment in some universal screening programs for secondary students due to efficiency of administration, future research should specifically investigate the reliability of CBM maze using generalizability theory in these grades, as well as the extent to which results are similar in students with disabilities. In addition, we explicitly evaluated fidelity of scoring procedures through calculations of inter-scorer agreement; however, we did not evaluate fidelity of administration procedures. This concern is partially mitigated by the prior formal training and experience the doctoral students had in CBM, but additional information on fidelity of administration would increase confidence in the findings.

Conclusions

Despite suggestions that CBM maze can be used as a time-efficient replacement for ORF assessments in universal screening (Fuchs & Fuchs, 1992; Graney et al., 2010), the results of this study indicate that the typical practice of administering one CBM maze probe per occasion is likely insufficient for screening and progress monitoring. Administration of two or three (depending on grade) 3-minute probes is necessary to reliably rank order students, which is

important to consider when identifying students in need of more intensive intervention and support based on performance relative to peers. Given that administration of two to three CBM maze probes likely is feasible for teachers and school psychologists, this study suggests that existing CBM maze measures may be adequate when group administration is necessary or desirable for universal screening so long as multiple probes are collected per occasion.

When the purpose of the CBM maze assessment is to make an absolute decision, such as identifying a specific student as at risk by comparing the student's score to a benchmark score or comparing a student's score on a later occasion to a score on a prior occasion, scores on multiple CBM maze probes (averages of two to four probes depending on grade) should be considered. If CBM maze is used in progress monitoring, moving averages (i.e., averages of subsets of the full data series) could be the basis of comparison. For example, if the average of three probes later in the series of scores for a specific student is greater than the average of three scores at the beginning of the series, with the number of probes selected depending on the student's grade, it would be appropriate to infer that the student's reading skills have improved. It is possible that future refinement through field testing of CBM maze probes to reduce discrepancies in difficulty would result in fewer probes needing to be administered for reliable decisions. With existing CBM maze probes, however, the results of this study suggest that more than one CBM maze probe needs to be administered for reliable scores in research and reliable decisions in practice.

References

Ardoin, S. P., Witt, J. C., Suldo, S. M., Connell, J. E., Koenig, J. L., Resetar, J. L., et al. (2004). Examining the incremental benefits of administering a maze and three versus one curriculum-based measurement reading probes when conducting universal screening. *School Psychology Review, 33*, 218-233.

Ardoin, S. P., Williams, J. C., Christ, T. J., Klubnik, C., & Wellborn, C. (2010). Examining readability estimates' predictions of students' oral reading rate: Spache, Lexile, and Forcast. *School Psychology Review, 39*, 277-285.

Bates, D., Maechler, M., & Bolker, B. (2011). *lme4: Linear mixed-effects models using S4 classes*. R package (Version 0.999375-39). http://CRAN.R-project.org/package=lme4

Betts, J., Pickart, M., & Heistad, D. (2009). An investigation of the psychometric evidence of CBM-R passage equivalence: Utility of readability statistics and equating for alternate forms. *Journal of School Psychology, 47*, 1-17. doi: 10.1016/j.jsp.2008.09.001

Brown-Chidsey, R., Davis, L., & Maya, C. (2003). Sources of variance in curriculum-based measures of silent reading. *Psychology in the Schools, 40*, 363-377. doi:10.1002/pits.10095

Christ, T. J., & Ardoin, S. P. (2009). Curriculum-based measurement of oral reading: Passage equivalence and probe-set development. *Journal of School Psychology, 47*, 55-75. doi:10.1016/j.jsp.2008.09.004

Christ, T. J., Johnson-Gros, K. N., & Hintze, J. M. (2005). An examination of alternate assessment durations when assessing multiple-skill computational fluency: The generalizability and dependability of curriculum-based outcomes within the context of educational decisions. *Psychology in the Schools, 42*, 615-622. doi: 10.1002/pits.20107

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

Deno, S. L., Reschly, A. L., Lembke, E. S., Magnusson, D., Callender, S. A., Windram, H., & Stachel, N. (2009). Developing a school-wide progress-monitoring system. *Psychology in the Schools, 46*, 44-55. doi:10.1002/pits.20353

Edformation. (2003). AIMSweb MAZE-comprehension curriculum-based measures. Retrieved August 18, 2010 from http://aimsweb.com/.

Espin, C., Wallace, T., Lembke, E., Campbell, H., & Long, J. D. (2010). Creating a progress-monitoring system in reading for middle-school students: Tracking progress toward meeting high-stakes standards. *Learning Disabilities Research & Practice, 25*, 60-75. doi:10.1111/j.1540-5826.2010.00304.x

Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21*, 45-58.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*, 20-28. doi:10.1177/074193258800900206

Graney, S. B., Martínez, R. S., Missall, K. N., & Aricak, O. T. (2010). Universal screening of reading in late elementary school: R-CBM versus CBM Maze. *Remedial and Special Education, 31*, 368-377. doi: 10.1177/0741932509338371

Hintze, J. M., Owen, S. V., Shapiro, E. S., & Daly, E. J. (2000). Generalizability of oral reading fluency measures: Application of G theory to curriculum-based measurement. *School Psychology Quarterly, 15*, 52-68. doi:10.1037/h0088778

Howe, K. B. & Shinn, M. M. (2002). *Standardized reading assessment passages (RAPs) for use in general outcome measurement: A manual describing development and technical features.* Retrieved July 30, 2010, from https://aimsweb.pearson.com

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Oxford, England: Addison-Wesley.

Marcoulides, G. A. (1990). An alternative method for estimating variance components in generalizability theory. *Psychological Reports, 66,* 379-386. doi: 10.2466/pr0.66.2.379-386

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Poncy, B. C., Skinner, C. H., & Axtell, P. K. (2005). An investigation of the reliability and standard error of measurement of words read correctly per minute using curriculum-based measurement. *Journal of Psychoeducational Assessment, 23,* 326-338. doi:10.1177/073428290502300403

R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

Shin, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *The Journal of Special Education, 34,* 164-172. doi:10.1177/002246690003400305

Shinn, M. R., & Shinn, M. M. (2002). *Administration and scoring of reading maze for use in general outcome measurement* Retrieved from http://www.aimsweb.com

Table 1

Correct Maze Choices by Probe, Grade, and Assessment Duration

		1 Minute		2 Minutes		3 Minutes	
Grade	Passage	*M*	*SD*	*M*	*SD*	*M*	*SD*
Third	1	4.32	2.36	8.65	3.75	13.41	5.66
	2	4.57	1.95	7.79	3.63	11.91	5.70
	3	3.90	2.70	9.06	4.25	13.90	5.86
	4	6.06	2.30	11.24	3.85	16.50	5.44
	5	4.63	2.25	9.21	3.97	13.79	5.82
	6	3.82	2.27	8.30	3.56	12.76	5.05
	7	4.58	2.28	9.55	3.80	14.71	5.50
	8	4.26	2.70	9.41	4.27	13.79	5.85
	9	5.75	2.54	11.76	4.25	16.93	5.76
Fourth	1	5.07	1.61	9.19	3.20	14.91	4.66
	2	3.25	1.45	5.97	2.72	9.73	4.43
	3	5.45	2.34	9.78	3.97	14.65	6.15
	4	3.51	1.64	7.26	2.97	11.71	4.83
	5	5.05	2.20	10.33	3.77	15.22	5.63
	6	5.76	2.25	10.79	4.18	16.27	5.97
	7	5.50	2.90	10.76	4.78	15.96	6.56
	8	6.23	2.91	11.90	4.79	17.39	6.60
	9	3.79	2.14	8.84	4.19	14.09	5.80
Fifth	1	6.79	3.55	12.49	6.18	18.35	8.91
	2	6.12	2.49	12.51	4.77	18.98	7.06
	3	8.71	3.43	16.06	5.83	23.65	8.75
	4	7.97	2.79	14.76	5.83	22.23	8.15
	5	6.97	2.38	12.15	4.39	17.80	6.56
	6	7.00	3.56	14.99	7.17	23.45	10.21
	7	6.67	3.24	13.20	6.27	19.62	8.77
	8	8.13	2.79	16.19	5.92	23.37	8.14
	9	4.89	2.49	10.97	4.99	17.51	7.29

Table 2

Variance Components and Percent of Total Variance by Grade and Assessment Durations

		1 Minute		2 Minutes		3 Minutes	
Grade	Facet	Estimate (*SE*)	%Total	Estimate (*SE*)	%Total	Estimate (*SE*)	%Total
3	*Person*	2.46 (.46)	40	8.28 (1.44)	50	19.07 (3.22)	57
	Probe	0.52 (.28)	8	1.56 (0.83)	9	2.43 (1.30)	7
	Residual	3.21 (.20)	52	6.89 (0.42)	41	11.77 (0.72)	36
4	*Person*	2.48 (.44)	41	8.49 (1.44)	46	19.90 (3.30)	53
	Probe	1.15 (.60)	19	3.40 (1.75)	18	5.50 (2.84)	14
	Residual	2.46 (.15)	40	6.68 (0.41)	36	12.38 (0.76)	33
5	*Person*	5.28 (.82)	52	23.50 (3.47)	64	51.82 (7.52)	69
	Probe	1.24 (.65)	12	3.56 (1.85)	10	7.26 (3.75)	9
	Residual	3.67 (.23)	36	9.41 (0.59)	26	16.17 (1.02)	22

Table 3

Standard Error of Measurement by Grade, Decision Type, and Number of Probes

		Number of Probes				
Grade	Decision	1	2	3	4	5
3	Relative	3.43	2.43	1.98	1.72	1.53
	Absolute	3.77	2.66	2.18	1.88	1.69
4	Relative	3.52	2.49	2.03	1.76	1.57
	Absolute	4.23	2.99	2.44	2.11	1.89
5	Relative	4.02	2.84	2.32	2.01	1.80
	Absolute	4.84	3.42	2.79	2.42	2.16

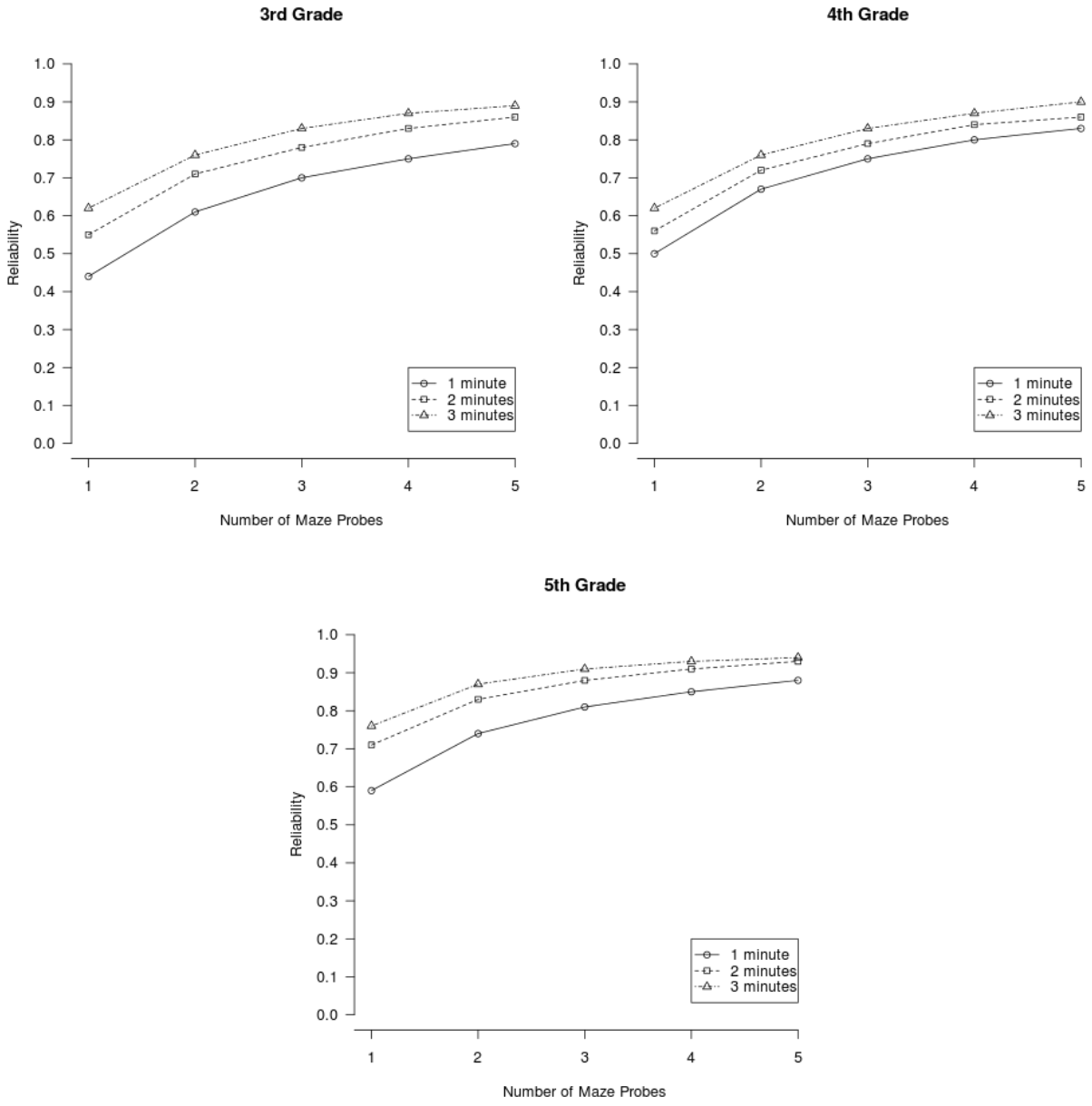

Figure 1. Reliability estimates for relative decisions by grade, number of probes, and assessment duration.

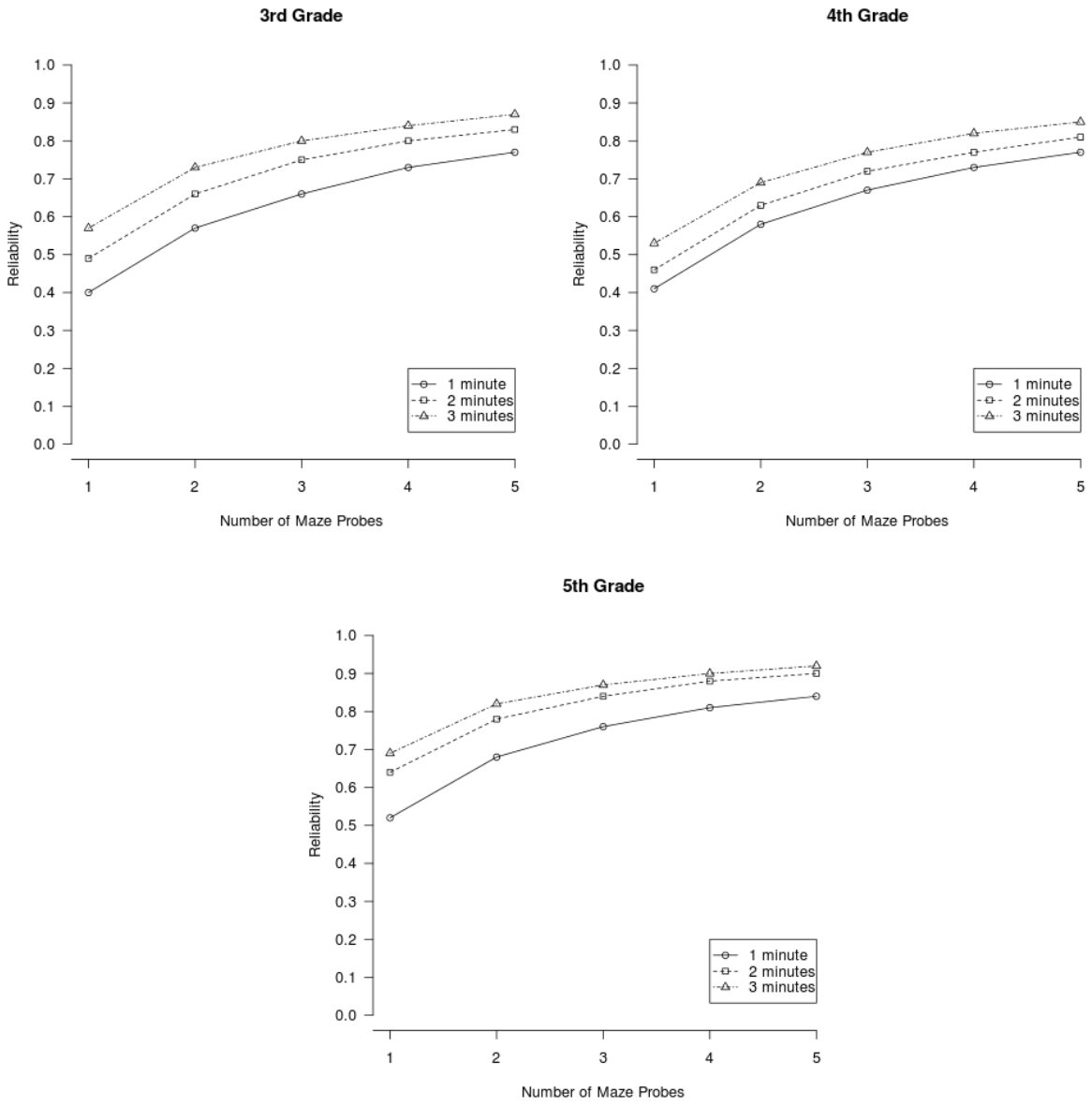

Figure 2. Reliability estimates for absolute decisions by grade, number of probes, and assessment duration.