ED 451 258 TM 032 486

AUTHOR	Lee, Guemin; Lewis, Daniel M.
TITLE	A Generalizability Theory Approach toward Estimating Standard Errors of Cutscores Set Using the Bookmark Standard Setting Procedure.
PUB DATE	2001-04-11
NOTE	31p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Seattle, WA, April 11-13, 2001).
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Cutting Scores; Elementary School Students; Elementary Secondary Education; *Error of Measurement; *Estimation (Mathematics); *Generalizability Theory; Item Response Theory; Secondary School Students; Standards; State Programs; Testing Programs
IDENTIFIERS	*Standard Setting

ABSTRACT

The Bookmark Standard Setting Procedure (Lewis, Mitzel, and Green, 1996) is an item-response-theory-based standard setting method that has been widely implemented by state testing programs. The primary purposes of this study were to: (1) estimate standard errors for cutscores that result from Bookmark standard settings under a generalizability theory model; and (2) to investigate the effects of different universes of generalization and several error sources on the standard errors. Data were obtained from a Bookmark standard setting conducted in grades 5, 8, and 11 mathematics for a large-scale assessment. The findings show that different patterns of error scores are found for different cut scores. Therefore, it seems necessary to estimate separate variance components for each cut score and to apply them to estimate the corresponding standard error. In addition, different universes of generalization produce different standard error estimates. As a result, policymakers should consider which universe is appropriate for the proposed use of the cutscores. There were also nonnegligible effects of participants and groups among several error sources. Increasing the number of small groups tended to be more efficient than increasing the number of participants per group given a fixed number of participants to decrease the standard errors. (Contains 5 tables, 7 figures, and 10 references.) (Author/SLD)

A Generalizability Theory Approach toward Estimating Standard Errors of Cutscores Set Using the Bookmark Standard Setting Procedure

Guemin Lee
Daniel M. Lewis
CTB/McGraw-Hill

Paper Presented at the Annual Meeting
of the National Council on Measurement in Education
Seattle, WA
April 11, 2001

Abstract

The Bookmark Standard Setting Procedure (Lewis, Mitzel, & Green, 1996) is an IRT-based standard setting method that has become widely implemented by state testing programs. The primary purposes of this study are to (a) estimate standard errors for cutscores that result from Bookmark standard settings under a generalizability theory model and (b) to investigate the effects of different universes of generalization and several error sources on the standard errors. The findings produced several notable results. First, different patterns of variance component estimates were found for different cut scores. Therefore, it seems necessary to estimate separate variance components for each cut score and to apply them to estimate the corresponding standard error. Second, different universes of generalization produced different standard error estimates and thus, policy makers should consider which universe is appropriate for the proposed use of the cutscores. Third, there were non-negligible effects of participants and groups among several error sources. Increasing the number of small groups tended to be more efficient than increasing the number of participants per group given a fixed number of participants to decrease the standard errors.

A Generalizability Theory Approach toward Estimating Standard Errors of Cutscores Set Using the Bookmark Standard Setting Procedure

Setting performance standards has become commonplace due to the standards-based education reform movement, Title I requirements, and public demands for accountability. Various methods have been developed to set cutscores for an assessment used to measure students' progress towards performance standards; a comprehensive review is summarized in Kane (1994), Jaeger (1989), Berk (1986), and Shepard (1980). The Bookmark Standard Setting Procedure (Lewis, Mitzel, & Green, 1996; Lewis, Green, Mitzel, Baum, & Patz, 1998) is an IRT-based standard setting procedure that was first developed in 1996. Since its inception, the Bookmark Procedure has become widely implemented—18 states have set cut scores for large-scale assessment programs using the procedure since 1996.

It is important that policy makers consider the uncertainty associated with the recommended cutscores that result from a formal standard setting prior to adopting operational cutscores. That is, sponsoring agencies and policy-making bodies should take the standard error associated with cutscores into account when considering modifying recommended cutscores (Lewis, 1997). However, data arising in standard setting contexts have complex dependency structures and reflect many sources of error. For example, one could consider the error associated with the selection of participants or the error associated with the selection of items considered by the participants when making their judgments. There are relatively few studies in the literature that examine issues related to standard errors associated with standard setting participants' recommended cutscores. Brennan and Lockwood (1980) and Kane and Wilson (1984) studied this topic, however they focused on the Angoff (1971) and Nedelsky (1954) procedures.

The present study is intended to result in a new approach to estimate the standard error of cutscores set using the Bookmark standard setting procedure under a generalizability theory model. The three primary objectives of the present study were:

1. to explore procedures to estimate standard errors of cutscores set using the Bookmark standard setting procedure under a generalizability theory model,
2. to determine the influence of different conceptualizations about the universe of generalization on the standard errors of cutscores set using the Bookmark standard setting procedure, and
3. to investigate the relative effects of standard setting participants, small groups, rounds, and forms, which constitute facets for the Bookmark standard setting method.

Brief Overview of the Bookmark Standard Setting Procedure

Item response theory (IRT) provides a framework that simultaneously characterizes the proficiency of examinees and the difficulty of test items. Just as it is possible to order examinees by estimated proficiency, IRT enables items to be ordered by the proficiency needed to have a specified probability of success. The facility to order items on the IRT proficiency scale is fundamental to the Bookmark procedure. Ordered item booklets are created by placing the items/score points in ascending order on the IRT scale. Each selected-response item is located on the IRT ability scale at the point at which a student would have a 0.67 probability of success, with guessing factored out. Each constructed-response score point is located at the point at which a student would have a 0.67 probability of achieving that score point or higher (Lewis, et. al. 1998). The cutscore for a given performance level, for example, "Proficient", can be identified by placing a bookmark between two items in the ordered item booklet such that from the participant's perspective, the items preceding the bookmark represent content that all proficient students should master (with mastery defined as having at least a 2/3 likelihood of success on the item/score point). By placing the bookmark at the furthest most item for which this is true, a unique location on the ability scale can be estimated as the cutscore. For a detailed account of the Bookmark Procedure, see Lewis, et. al. (1998).

A Generalizability Theory Model

The observed score of a standard setting participant for this study can be partitioned into several effects under a $(p:g:f)\times r$ univariate generalizability study (G-study) design: participants (*p*) nested within small groups (*g*) nested within forms (*f*) crossed with rounds (*r*). The linear model is expressed as

$$X_{pgfr} = \mu \sim + \mu_{p:g:f} \sim + \mu_{g:f} \sim + \mu_{f} \sim + \mu_{r} \sim + \mu_{rf} \sim + \mu_{rg:f} \sim + \mu_{rp:g:f,e} \sim , \tag{1}$$

where the terms on the right-hand side are the grand mean, participant in group nested within form effect, group nested within form effect, form effect, round effect, round by form interaction effect, round by group nested within form interaction effect, round by participant in group nested within form interaction effect confounded with unexplained sources of error, respectively.

A G-study is done to determine how well the scores can be used for multiple situations, and involves estimating variance components that might in turn be used in a decision study (D-study). A D-study is a study conducted for computing reliability-like coefficients and standard errors of measurement and/or for the purpose of determining the most efficient measurement procedures for a given situation. The most important D-study consideration is the specification of a universe of generalization, to which a decision-maker wants to generalize a score with a particular measurement procedure (Brennan, 1992; 2000). The standard errors (SE) for cutscores that result from a standard setting are dependent upon the investigator's or policy maker's specifications about facets over which universe of generalization is to be considered. In the current study, four universes of generalization are considered, and associated formulae for estimating SEs for the Bookmark procedure are provided.

Suppose an investigator, David, decided that the participants, small groups, forms (test forms), and rounds constitute facets in the universe of generalization, and each replication involves different sets of participants, small groups, forms, and rounds. In this case, David's universe can be considered an infinite universe of generalization and each facet is treated as random. The SE for the Bookmark procedure for David's universe of generalization can be estimated by

$$SE = \sqrt{\frac{\hat{\sigma}^2(p:g:f)}{n'_p n'_g n'_f} + \frac{\hat{\sigma}^2(g:f)}{n'_g n'_f} + \frac{\hat{\sigma}^2(f)}{n'_f} + \frac{\hat{\sigma}^2(r)}{n'_r} + \frac{\hat{\sigma}^2(rf)}{n'_r n'_f} + \frac{\hat{\sigma}^2(rg:f)}{n'_r n'_g n'_f} + \frac{\hat{\sigma}^2(rp:g:f)}{n'_r n'_p n'_g n'_f}} \tag{2}$$

where the estimates of variance components from a G-study are:

$\hat{\sigma}^2(p:g:f) =$ the estimate of variance among participants in small groups within forms;

$\hat{\sigma}^2(g:f) =$ the estimate of variance among small groups in forms;

$\hat{\sigma}^2(f) =$ the estimate of variance of forms;

$\hat{\sigma}^2(r) =$ the estimate of variance of rounds;

$\hat{\sigma}^2(rf) =$ the estimate of variance for interactions of rounds and forms;

$\hat{\sigma}^2(rg:f) =$ the estimate of variance for interaction of rounds and small groups in forms;

$\hat{\sigma}^2(rp:g:f) =$ the estimate of variance for interactions of rounds and participants in small groups within forms; and n'_p, n'_g, n'_f, and n'_r represent sample size for participants, small groups, forms, and rounds, respectively, in a D-study.

Another investigator, Karla, might not be interested in generalizing scores over forms. That is, she would have a conceptualization about replications composed of different participants, different groups, and different rounds with the same form. Thus, Karla's universe of generalization is "restricted" in that it contains a fixed facet (forms). However, this does not mean that David's universe is better than Karla's universe. Two investigators merely have different conceptualizations about the universe of generalization. The SE for Karla's universe of generalization for the standard setting procedure is

$$SE = \sqrt{\frac{\hat{\sigma}^2(p:g:f)}{n'_p n'_g n'_f} + \frac{\hat{\sigma}^2(g:f)}{n'_g n'_f} + \frac{\hat{\sigma}^2(r)}{n'_r} + \frac{\hat{\sigma}^2(rf)}{n'_r n'_f} + \frac{\hat{\sigma}^2(rg:f)}{n'_r n'_g n'_f} + \frac{\hat{\sigma}^2(rp:g:f)}{n'_r n'_p n'_g n'_f}}\,. \qquad (3)$$

Suppose another investigator, Robert, is not interested in generalizing scores over rounds. He would have a conceptualization about replications composed of different participants, different groups, different forms, and the same rounds. Thus, Robert's universe of generalization is "restricted" compared to David's because it contains a fixed round facet. The SE for Robert's universe of generalization for the standard setting procedure is

$$SE = \sqrt{\frac{\hat{\sigma}^2(p:g:f)}{n'_p n'_g n'_f} + \frac{\hat{\sigma}^2(g:f)}{n'_g n'_f} + \frac{\hat{\sigma}^2(f)}{n'_f} + \frac{\hat{\sigma}^2(rf)}{n'_r n'_f} + \frac{\hat{\sigma}^2(rg:f)}{n'_r n'_g n'_f} + \frac{\hat{\sigma}^2(rp:g:f)}{n'_r n'_p n'_g n'_f}} \,. \qquad (4)$$

Another investigator, Karen, might be interested in generalizing scores over only participants and small groups. That is, his conceptualization about replications is composed of only different participants and different small groups. Thus, Karen's universe of generalization is the most "restricted." The SE for Karen's universe of generalization can be estimated by

$$SE = \sqrt{\frac{\hat{\sigma}^2(p:g:f)}{n'_p n'_g n'_f} + \frac{\hat{\sigma}^2(g:f)}{n'_g n'_f} + \frac{\hat{\sigma}^2(rg:f)}{n'_r n'_g n'_f} + \frac{\hat{\sigma}^2(rp:g:f)}{n'_r n'_p n'_g n'_f}} \,. \qquad (5)$$

Method

Data Sources

The data used in this study were obtained from a Bookmark standard setting conducted in grades 5, 8, and 11 mathematics for a large-scale assessment. Three cut scores were set in each grade to define four performance levels: Level 1, Level 2, Level 3, and Level 4, with Level 1 being the highest achieving group. For each grade, alternate forms of the ordered item booklets were constructed by selecting items from various forms of the assessment, which were administered in an operational assessment. The two forms were constructed to be as parallel as possible in terms of objective structure and difficulty.

The standard setting participants for each grade and content area were divided into four small groups that worked independently through several rounds of judgments. Two of the four small groups worked with Form A of the ordered item booklet and the other two small groups used Form B. Prior to the first round of judgments, participants studied the ordered item booklets within their small groups, and discussed what each item measured and why each item was more difficult than the preceding items in the booklet. Following discussion, participants made individual and independent Round 1 judgments, that is they placed bookmarks that indicated the items that reflected content the participant expected students in the associated performance level to know and be able to do. In Round 2, participants discussed their Round 1 judgments within each small group and followed discussion with individual Round 2 judgments. In Round 3, participants using the same form of the ordered item booklet considered the estimated percent of students in each performance level based on their current judgments, discussed their Round 2 judgments, and concluded with individual Round 3 judgments. The fourth and final round consisted of discussion among all participants within a grade/content area followed by individual judgments. The response vectors obtained from the first three rounds of judgments were used as data sources in the current study.

Analyses

The computer application programs, GENOVA (Crick & Brennan, 1983) and urGENOVA (Brennan, 1999) were used in this study for estimating variance components for the Bookmark standard setting method using a $(p:g:f)\times r$ univariate generalizability G-study design, participants (*p*) nested within small groups (*g*) nested within forms (*f*) crossed with rounds (*r*). Because the number of participants per small group varied in the grade 8 data, the conditions for a balanced design were not met for this grade. Consequently, ANOVA-like procedures were used with urGENOVA to estimate variance components. With variance component estimates from the G-studies, SEs for the Bookmark cutscores were estimated using Equations 2 through 5. To determine the practical influence of standard setting participants, small groups, and rounds, D-studies were conducted with varying numbers of participants, groups, and rounds.

Results and Implications

Variance component estimates for the *(p:g:f)×r* G-study design for the Bookmark standard setting method are presented in Table 1. In each grade, three sets of estimates are provided for three cutscores, separately, along with the average of the three estimates over the three cutscores for each of variance components. The 'Proportion' column represents the percentage of total variance attributable to each variance component.

Insert Table 1 About Here

The largest variance component was the rounds by participants interaction term in small groups within forms, $\hat{\sigma}^2(rp:g:f,e)$. This term explained about 50% of total score variation on the average. Because this variance component includes variance components due to unexplained sources of error as well as a rounds by participants interaction effect, it is not surprising to find a relatively large magnitude for this term. Relatively small percents of the total variance were accounted for by participants in groups within forms component, $\hat{\sigma}^2(p:g:f)$: 3.9% for grade 5, 6.1% for grade 8, and 5.8% for grade 11.

Somewhat different patterns for variance component estimates were observed across grades. For example, the variance component estimate for groups within forms, $\hat{\sigma}^2(g:f)$, accounted for 3.9%, 7.7%, and 26.1% of the total score variance for grades 5, 8, and 11, respectively. In contrast, the rounds by small groups interaction term, $\hat{\sigma}^2(rg:f)$, accounted for 28.9%, 3.8%, and 0.8% of the total variance for grades 5, 8, and 11, respectively. Relatively large form, round, and form by round interaction variance component estimates were found in grade 8 compared to grades 5 and 11.

In addition to the different patterns for variance component estimates across the three grades, different variance component structures were also found over the three cutscores. For example, in grade 5 Mathematics, 12.4%, 9.8%, and 0.0% of the total variance were accounted for by groups within forms for cutscores 1, 2, and 3, respectively. Variance component for rounds by small groups interaction within forms explained 2.2% of total variance for cutscore 1, but it explained 25.2% for cutscore 2 and 34.3% for cutscore 3. The rounds by forms interaction term was 21.3% for cutscore 2, but it was just 3.0% for cutscore 1 and 0.0% for cutscore 3. The proportion of variance component for rounds took 0.0% for cutscore 1 and 2, and 12.0% for cutscore 3 among total score variance. Thus, it seems necessary to estimate the variance components for each cutscore separately.

The SEs for the Bookmark cutscores for four different universes of generalization are presented in Table 2 assuming similar measurement specifications in the G-study. That is, each D-study incorporated 1 form, 2 small groups per form, 6 participants per small group, and 3 rounds. Variance component estimates for each of three cutscores and average estimates were used to compute the SEs using Equations 2 through 5. As would be expected, Universe Type 1 produced the largest SE because it had a broader definition about the universe of generalization than did the others. Universe Type 4 was most restricted and, consequently, produced the smallest SE. Universe Type 2 (forms-fixed; rounds-random) and Universe Type 3 (forms-random; rounds-fixed) provided similar SEs across the three grades though the Universe Type 2 had a somewhat smaller SE. In general, relatively smaller SEs were reported for cutscore 1 than for cutscores 2 and 3 except for grade 8 Mathematics. The SE for cutscore 2 in grade 8 was similar to the SE for cutscore 1.

Insert Table 2 About Here

The four universe types produced somewhat different SEs, especially for cutscore 2 in grade 5 and for cutscore 3 in grade 8. For example, in grade 8, the SE for cutscore 3 for Universe Type 1 was 10.73, but it was just 2.46 for Universe Type 4. This large difference was likely to be related to the variance component estimates presented in Table 1. The variance components for forms, rounds, and rounds by forms interaction terms for cutscore 3 in grade 8 were large relative to other cutscores in other grades. Thus, treating forms and/or rounds as 'fixed' or 'random' had a strong influence on the SEs. For cutscore 2 in grade 8, these variance components had zero estimates and the four universe types produced the same SEs. That is, different conceptualizations about the universe of generalization had a non-negligible influence on the magnitude of the SEs. The influence seemed related to the magnitudes of variance component associated with 'forms' and 'rounds'.

The influence of round effects on the SEs was investigated by conducting several D-studies. The SEs with varying number of rounds for grade 5 Mathematics are summarized in Table 3. We can observe a substantial influence of round effects in this table. As the number of rounds increased, the SE decreased. An interaction between rounds effects and universe types can be observed. That is, more distinct differences could be found, on average, in Universe Types 1, 2, and 3 than in Universe Type 4. Also, there was an interaction between round effects and cutscores. The round effects were more evident in cutscores 2 and 3 than in cutscore 1. Similar trends were found in grades 8 and 11, summarized in Tables 4 and 5, respectively.

Insert Table 3 About Here

Insert Table 4 About Here

Insert Table 5 About Here

One notable difference in the grade 8 trends was that the rounds had a relatively small influence on the magnitude of the SEs for cutscore 2 and/or Universe Type 4 compared to other cutscores and/or other universe types. This could be anticipated from relatively smaller variance component estimates related to 'rounds' for this specific cutscore in this grade. Very similar SEs were reported regardless of specific cutscores and/or universe types for grade 11.

Effects of small groups and participants on SEs for grade 5 using average variance component estimates are presented in Figure 1. The horizontal axis represents the number of participants per small group and the vertical axis represents the SEs for the Bookmark cutscores. The top line shows SE changes when two small groups are used. The middle and bottom lines represent SE changes for three and four small groups, respectively. For example, for two small groups with seven participants per small group, the overall level SEs for Universe Type 2 would be around 7. As would be expected, the SEs for Universe Type 1 was higher than those for other universe types, and Universe Type 4 produced the lowest SEs.

Insert Figure 1 About Here

As the number of participants within groups increased, the SEs decreased. Also, when the number of small groups increased, there was a non-negligible decrease in the SEs. Similar trends were found for grades 8 and 11, presented in Figures 2 and 3, respectively. Relatively small group effects were found in grade 8 but relatively large group effects were found in grade 11.

Insert Figure 2 About Here

Insert Figure 3 About Here

The effects due to small group and those due to participants are confounded. That is, given a point of horizontal axis (e.g., n'_p=7 in Figure 1), the SEs for two small groups (n'_g=2) and three small groups (n'_g=3) were based upon different total number of participants. The number of total participants was 14 when n'_g=2, but it was 21 when n'_g=3. Therefore, smaller SE can be expected for the case involving a larger number of participants. To investigate the relationship between small group and participant effects in a more distinct manner, the SEs for Bookmark cutscores given the same total number of participants are presented in Figure 4.

Insert Figure 4 About Here

A case with three groups and three participants per group is illustrated. For example, given 21 participants, the '3 groups' line represents the SE for three groups and seven participants per group. In contrast, the '3 participants' line represents the SE for three participants per group and seven groups. These results show that for a given number of participants, two different grouping strategies result in different SEs; increasing the number of small groups results in greater reliability than increasing the number of participants per small group. Similar trends were found for grades 8 and 11, and are presented in Figures 5 and 6, respectively.

Insert Figure 5 About Here

Insert Figure 6 About Here

Similar results were produced for each cutscore within each grade, except cutscore 3 in grade 8. For this special case, both the '3 Groups' and '3 Participants' lines produced very similar SEs in all universe types. That is, two different types of grouping strategies did not make any significant differences in SEs. If the total number of participants are the same, both strategies produced very similar SEs. The SEs for Bookmark cutscores given the same total number of participants for cutscore 3 in grade 8 are presented in Figure 7.

Insert Figure 7 About Here

This result can be explained by the variance component estimates presented in Table 1. If variance component terms included both 'small groups' and 'participants' facets, the D-study variance components for those would be divided by sample sizes for both 'small groups' and 'participants'. Therefore, those variance components would not be influenced by different grouping strategies if the total number of participants ($= n'_p \times n'_g$) were the same. The variance component terms that just include 'small groups' will be influenced by different strategies of grouping the same number of total participants. These variance components are $\hat{\sigma}^2(g:f)$ and $\hat{\sigma}^2(rg:f)$. The magnitudes for these two components were smaller in grade 8 cutscore 3 than other grades and cutscores.

The results of this study suggest the following implications:

First, different patterns of variance component estimates can be anticipated for different cutscores; thus, it is necessary to estimate variance components and SE for each cutscore. One SE can not be applied to different Bookmark cutscores without sacrificing accuracy.

Second, different conceptualization about the universe of generalization result in different SEs for Bookmark cutscores. Thus, policy makers or decision-making bodies should consider the uncertainty associated with a specific definition of the universe of generalization; this must be specified prior to estimating the SE.

Third, there are non-negligible effects of small groups and participants. Even though these effects are confounded, the results suggest that for a fixed number of participants, increasing the number of small groups will result in increased reliability of cutscores. Groups that implement the Bookmark standard setting procedure can use this result to determine the most efficient standard setting configuration.

Discussion

Two alternate forms were constructed for this study and the Bookmark procedure was implemented using both forms. However, this is not typical. In most cases, Bookmark standard settings are conducted using *a* single test form to set cutscores. However, if the comparison of performance over forms (or over years using different forms) were of interest, it is necessary to incorporate form and related interactions as sources of error. A strong component of this study is that it allows the assessment of these factors.

In the results reported in Tables 3, 4, and 5, round facet was assumed to be "random". However, this is not completely true in reality. By the discussion within small groups, standard setting participants can change their judgments. Therefore, their judgements would be biased and variance components might be inaccurately estimated. However, estimating variance components for rounds and related interaction effects under this random-round assumption provides valuable information in assessing the error of Bookmark cutscores.

We now discuss how this SE information might be applied in practice. A cutscore is used for making mastery decisions by comparing an examinee's score and the cutscore. If the examinee's score is greater than or equal to a specific cutscore, the examinee is assumed to have the skills associated with that level. The main interest should be given to the comparison of:

$$SS_p - CS_l, \tag{6}$$

where SS_p represents scale score for pupil p, and CS_l is the cutscore for performance level l. The difference should reflect the true difference that we are really interested in:

$$TSS_p - TCS_l, \tag{7}$$

where TSS_p represents true scale score for pupil p, and TCS_l is the true cutscore for performance level l. Then, Equation (6) is an estimator of Equation (7). The error can be defined by the difference between the estimator and the true state:

$$\Delta_p = (SS_p - CS_l) - (TSS_p - TCS_l) \tag{8}$$

Consequently, the error variance in estimating the difference score, $(TSS_p - TCS_l)$, is

$$\sigma^2(\Delta_p) = Var[(SS_p - CS_l) - (TSS_p - TCS_l)] \tag{9}$$

$$= Var[(SS_p - TSS_p) - (CS_l - TCS_l)] \tag{10}$$

$$= Var(SS_p - TSS_p) + Var(CS_l - TCS_l) - 2Cov[(SS_p - TSS_p), (CS_l - TCS_l)] \tag{11}$$

where $Var(SS_p - TSS_p)$ represents measurement error variance for pupil p and $Var(CS_l - TCS_l)$ is error variance for the performance level l cutscore. By the assumption of $Cov[(SS_p - TSS_p), (CS_l - TCS_l)] = 0$, the standard error can be estimated by

$$\hat{\sigma}(\Delta_p) = \sqrt{V\hat{a}r(SS_p - TSS_p) + V\hat{a}r(CS_l - TCS_l)}\,, \tag{12}$$

This standard error is broader than that associated with measurement error alone and can be used by policy makers in making mastery decisions. For example, students who score below the cutscore but do so within the standard error band may in fact hold the requisite skills. This might be considered when making high stakes decisions about students. By the same token, students who score above the cutscore but do so with the standard error band may not in fact hold requisite skills and might be considered candidates for remediation.

References

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education.

Berk, R.A. (1986). A consumer's guide to setting performances standards on criterion-referenced tests. Review of Educational Research, 56, 137-172.

Brennan, R.L. (1992). Elements of generalizability theory. Iowa City, IA: American College Testing.

Brennan, R.L. (2000, April). An essay on the history and future of reliability from the perspective of replications. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Brennan, R.L., & Lockwood, R.E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4, 219-240.

Crick, J.E., & Brennan, R.L. (1983). Manual for GENOVA: A GENeralized analysis Of VAriance system. (ACT Technical Bulletin No. 43). Iowa City, IA: American College Testing Program.

Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.), Educational measurement (3rd ed.). New York: American Council on Education and Macmillan.

Kane, M. (1994). Validating the performance standards associated with passing scores. Review of Educational Research, 64, 425-461.

Kane, M., & Wilson, J. (1984). Errors of measurement and standard setting in mastery testing. Applied Psychological Measurement, 8, 107-115.

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1998, April). The Bookmark Standard Setting Procedure: Methodology and Recent Implementations. Paper presented at the 1998 National Council for Measurement in Education annual meeting, San Diego, CA.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A Bookmark approach. In D.R. Green (Chair), IRT-based standard setting procedures utilizing behavioral anchoring. Symposium presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Lewis, D. M. (1997). Overview of the Standard Errors Associated with Standard Setting. Unpublished research paper.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Shepard, L. (1980). Standard setting, issues, and methods. Applied Psychological Measurement, 4, 447-467.

TABLE 1

Variance Component Estimates for the Random Effects $(p:g:f) \times r$ Generalizability Study Design for the Bookmark Standard Setting Procedures

Variance Component	Cut Score 1		Cut Score 2		Cut Score 3		Average	
	Estimate	Proportion	Estimate	Proportion	Estimate	Proportion	Estimate	Proportion
				Grade 5 Mathematics				
$\hat{\sigma}^2(p:g:f)$	6.7	6.4	15.9	4.8	23.0	3.1	15.2	3.9
$\hat{\sigma}^2(g:f)$	12.9	12.4	32.5	9.8	0.0	0.0	15.1	3.9
$\hat{\sigma}^2(f)$	11.0	10.6	43.0	13.0	2.5	0.3	18.8	4.8
$\hat{\sigma}^2(r)$	0.0	0.0	0.0	0.0	88.1	12.0	29.4	7.5
$\hat{\sigma}^2(rf)$	3.1	3.0	70.5	21.3	0.0	0.0	24.5	6.3
$\hat{\sigma}^2(rg:f)$	2.3	2.2	83.3	25.2	252.6	34.3	112.7	28.9
$\hat{\sigma}^2(rp:g:f,e)$	68.0	65.4	85.3	25.8	370.9	50.3	174.7	44.7
				Grade 8 Mathematics				
$\hat{\sigma}^2(p:g:f)$	17.6	15.5	0.0	0.0	14.9	4.3	10.8	6.1
$\hat{\sigma}^2(g:f)$	14.9	13.1	23.2	32.2	0.0	0.0	12.7	7.2
$\hat{\sigma}^2(f)$	0.0	0.0	0.0	0.0	67.3	19.5	22.4	12.7
$\hat{\sigma}^2(r)$	0.0	0.0	0.0	0.0	68.1	19.7	22.7	12.8
$\hat{\sigma}^2(rf)$	29.3	25.8	0.0	0.0	56.9	16.5	28.7	16.2
$\hat{\sigma}^2(rg:f)$	0.0	0.0	13.4	18.6	7.1	2.1	6.8	3.8
$\hat{\sigma}^2(rp:g:f,e)$	51.7	45.6	35.4	49.2	131.4	38.0	72.8	41.2
				Grade 11 Mathematics				
$\hat{\sigma}^2(p:g:f)$	24.7	28.2	52.3	16.7	10.6	0.9	29.2	5.8
$\hat{\sigma}^2(g:f)$	16.5	18.8	116.6	37.3	263.5	23.5	132.2	26.1
$\hat{\sigma}^2(f)$	20.4	23.3	65.9	21.1	11.3	1.0	32.5	6.4
$\hat{\sigma}^2(r)$	0.3	0.3	0.0	0.0	0.0	0.0	0.1	0.0
$\hat{\sigma}^2(rf)$	0.0	0.0	1.4	0.4	48.1	4.3	16.5	3.3
$\hat{\sigma}^2(rg:f)$	1.6	1.8	7.4	2.4	0.0	0.0	3.0	0.8
$\hat{\sigma}^2(rp:g:f,e)$	24.2	27.6	69.3	22.1	784.9	70.2	292.8	57.8

TABLE 2

Standard Errors for the Bookmark Cutscores for Four Different Universes of Generalization

	Standard Erorr			
Universe of Generalization	Cut Score 1	Cut Score 2	Cut Score 3	Average
		Grade 5 Mathematics		
Universe Type 1	4.62	10.02	9.28	8.32
Universe Type 2	3.21	7.57	9.15	7.10
Universe Type 3	4.62	10.02	7.71	7.72
Universe Type 4	3.04	5.82	7.37	5.70
		Grade 8 Mathematics		
Universe Type 1	4.49	3.85	10.73	7.07
Universe Type 2	4.49	3.85	6.91	5.25
Universe Type 3	4.49	3.85	9,61	6.51
Universe Type 4	3.22	3.85	2.46	3.23
		Grade 11 Mathematics		
Universe Type 1	5.63	11.50	13.48	10.73
Universe Type 2	3.37	8.14	13.06	9.09
Universe Type 3	5.63	11.50	13.48	10.73
Universe Type 4	3.35	8.11	12.43	8.78

Notes. Universe Type 1 = Form-Random, Round-Random, Universe Type 2 = Form-Fixed, Round-Random, Universe Type 3 = Form-Random, Round-Fixed, Universe Type 4 = Form-Fixed, Round-Fixed. Each decision study assumed 1 form, 2 small groups per form, 6 participants per small group, and 3 rounds.

TABLE 3
The Round Effect on Standard Errors for the Bookmark Cutscores
(Grade 5 Mathematics)

Number of Rounds	Standard Erorr			
In Decision Study	Cut Score 1	Cut Score 2	Cut Score 3	Average
		Universe Type 1		
1	5.28	13.41	15.80	12.35
2	4.79	10.96	11.27	9.49
3	4.62	10.02	9.28	8.32
4	4.53	9.51	8.11	7.67
		Universe Type 2		
1	4.11	11.70	15.72	11.56
2	3.46	8.79	11.16	8.44
3	3.21	7.57	9.15	7.10
4	3.08	6.88	7.95	6.33
		Universe Type 3		
1	5.28	13.41	12.71	11.10
2	4.79	10.96	9.11	8.68
3	4.62	10.02	7.54	7.71
4	4.53	9.51	6.61	7.18
		Universe Type 4		
1	3.72	8.14	12.61	8.93
2	3.22	6.48	8.97	6.66
3	3.05	5.82	7.37	5.70
4	2.95	5.46	6.42	5.15

Notes. Universe Type 1 = Form-Random, Round-Random, Universe Type 2 = Form-Fixed, Round-Random, Universe Type 3 = Form-Random, Round-Fixed, Universe Type 4 = Form-Fixed, Round-Fixed. Each decision study assumed 1 form, 2 small groups per form, 6 participants per small group, and 3 rounds.

TABLE 4
The Round Effect on Standard Errors for the Bookmark Cutscores
(Grade 8 Mathematics)

Number of Rounds In Decision Study	Standard Erorr			
	Cut Score 1	Cut Score 2	Cut Score 3	Average
		Universe Type 1		
1	6.52	4.61	14.42	9,52
2	5.07	4.05	11.76	7.76
3	4.49	3.85	10.73	7.07
4	4.16	3.74	10.17	6.70
		Universe Type 2		
1	6.52	4.61	11.86	8.26
2	5.07	4.05	8.43	6.14
3	4.49	3.85	6.91	5.25
4	4.16	3.74	6.01	4.74
		Universe Type 3		
1	6.52	4.61	11.83	8.24
2	5.07	4.05	10.21	6.99
3	4.49	3.85	9.61	6.51
4	4.16	3.74	9.29	6.26
		Universe Type 4		
1	3.64	4.61	3.97	4.09
2	3.33	4.05	2.91	3.46
3	3.22	3.85	2.46	3.23
4	3.16	3.74	2.21	3.10

Notes. Universe Type 1 = Form-Random, Round-Random, Universe Type 2 = Form-Fixed, Round-Random, Universe Type 3 = Form-Random, Round-Fixed, Universe Type 4 = Form-Fixed, Round-Fixed. Each decision study assumed 1 form, 2 small groups per form, 6 participants per small group, and 3 rounds.

TABLE 5
The Round Effect on Standard Errors for the Bookmark Cutscores
(Grade 11 Mathematics)

Number of Rounds In Decision Study	Standard Erorr			
	Cut Score 1	Cut Score 2	Cut Score 3	Average
		Universe Type 1		
1	5.82	11.81	16.04	11.98
2	5.68	11.58	14.17	11.06
3	5.63	11.50	13.48	10.73
4	5.61	11.46	13.13	10.57
		Universe Type 2		
1	3.66	8.58	15.69	10.54
2	3.44	8.25	13.76	9.48
3	3.37	8.14	13.06	9.09
4	3.33	8.09	12.69	8.90
		Universe Type 3		
1	5.79	11.81	16.04	11.98
2	5.67	11.58	14.17	11.06
3	5.63	11.50	13.48	10.73
4	5.60	11.46	13.13	10.57
		Universe Type 4		
1	3.62	8.49	14.07	9.72
2	3.42	8.21	12.86	9.03
3	3.35	8.11	12.43	8.78
4	3.32	8.06	12.21	8.66

Notes. Universe Type 1 = Form-Random, Round-Random, Universe Type 2 = Form-Fixed, Round-Random, Universe Type 3 = Form-Random, Round-Fixed, Universe Type 4 = Form-Fixed, Round-Fixed. Each decision study assumed 1 form, 2 small groups per form, 6 participants per small group, and 3 rounds.

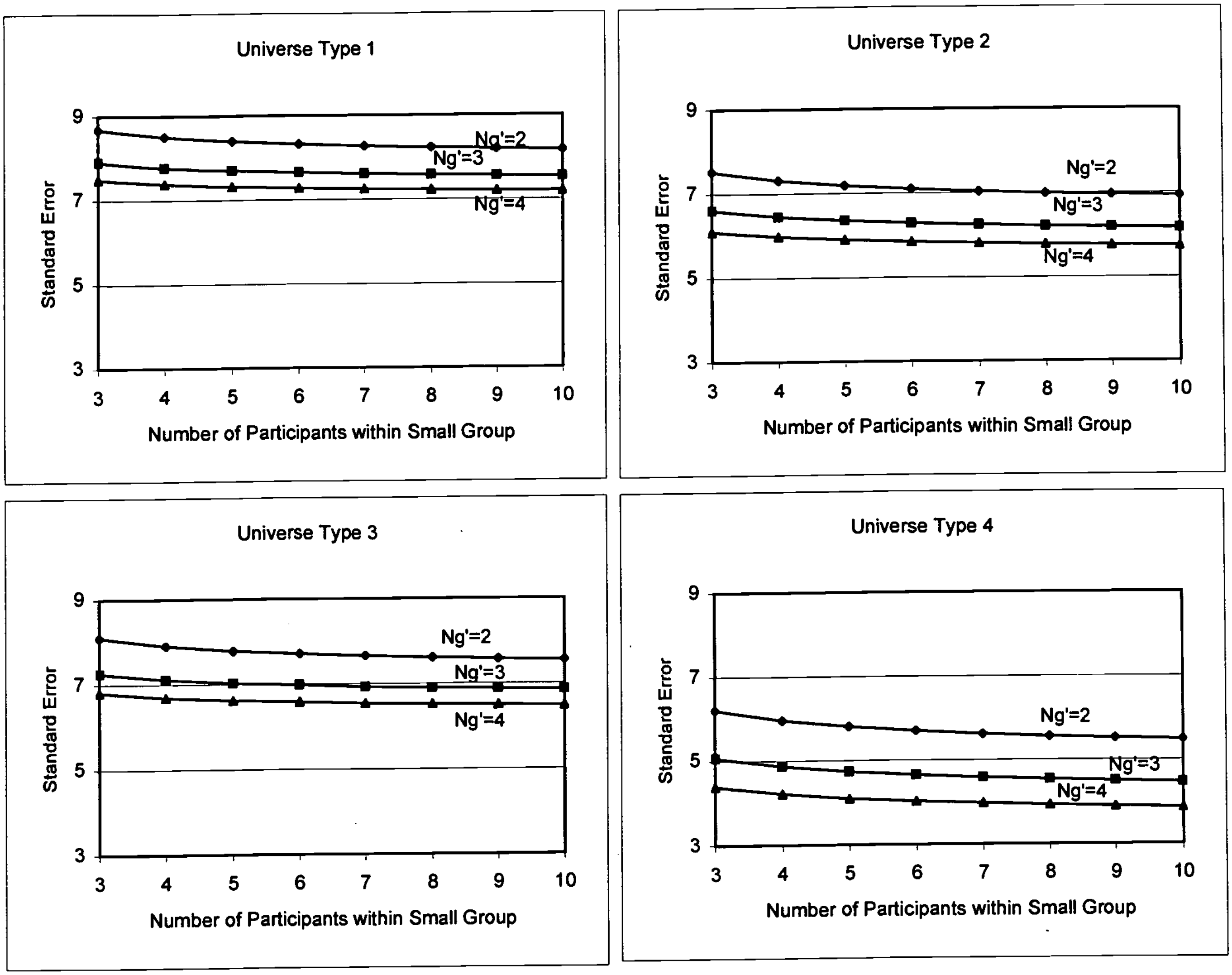


Figure 1 The effects of small groups and participants within small group on standard errors for the Bookmark Cutscores (Grade 5).

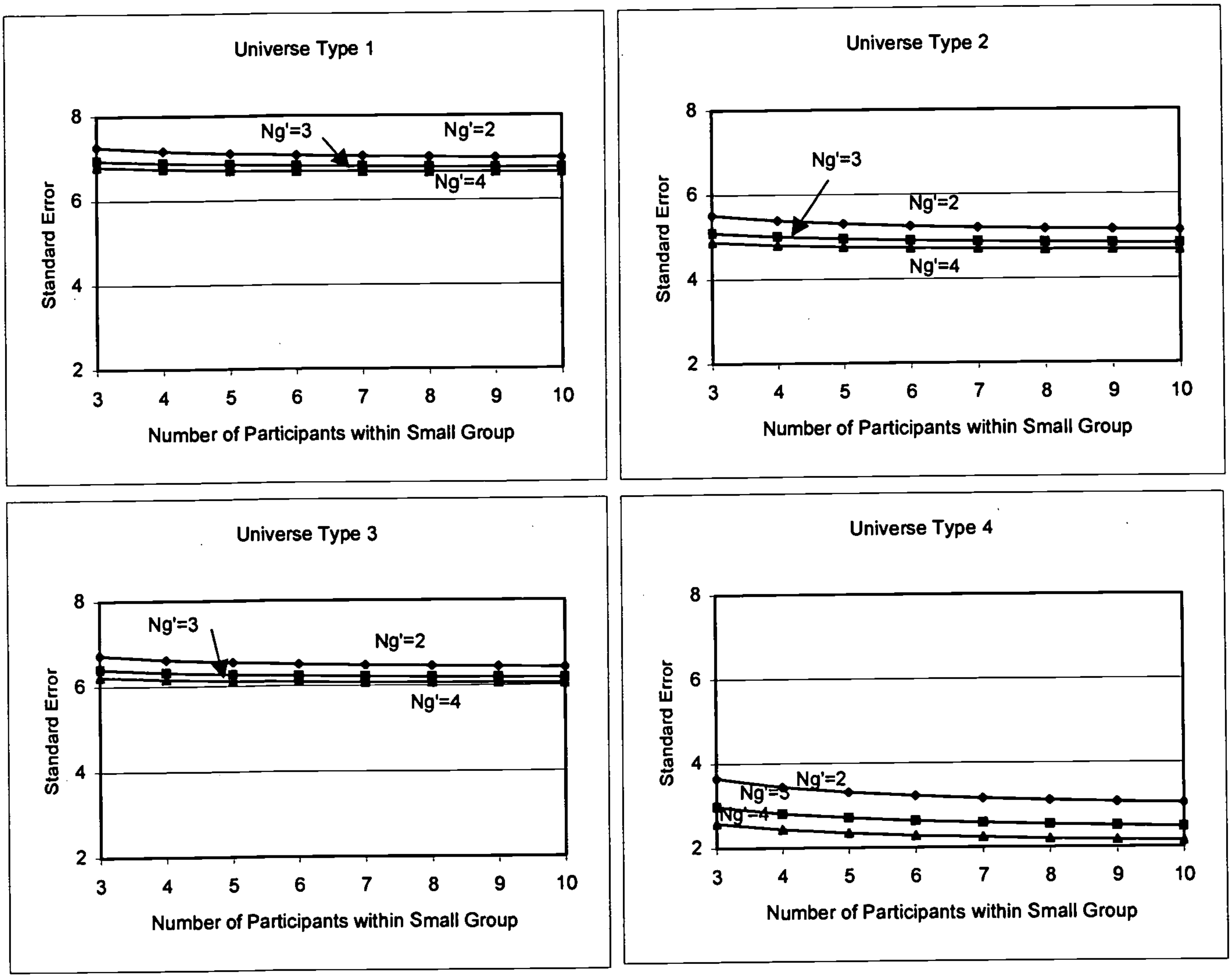


Figure 2. The effects of small groups and participants within small group on standard errors for the Bookmark Cutscores (Grade 8).

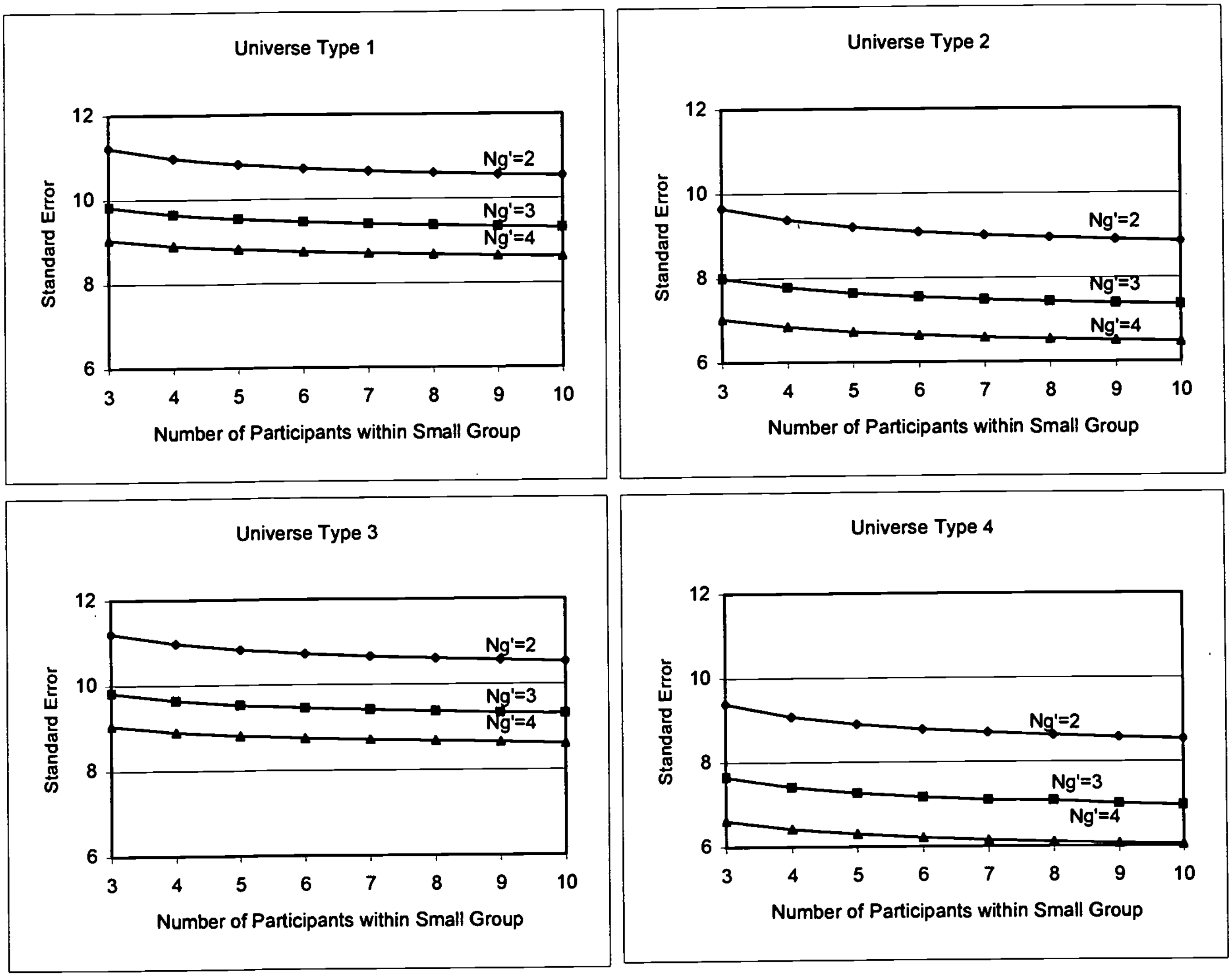


Figure 3. The effects of small groups and participants within small group on standard errors for the Bookmark Cutscores (Grade 11).

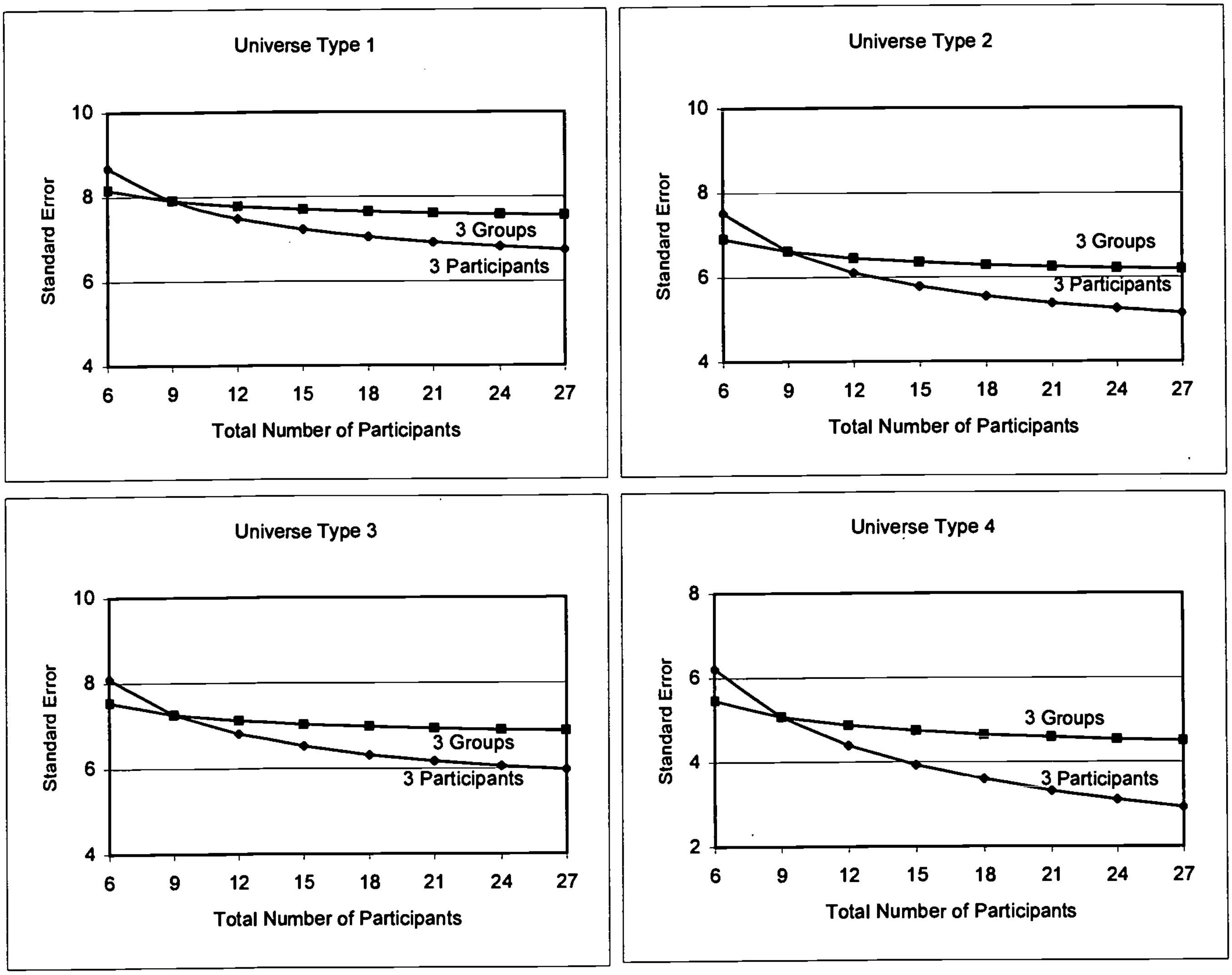


Figure 4. The effects of small groups and participants within small group on standard errors for the Bookmark Cutscores for Given Total Number of Panels (Grade 5).

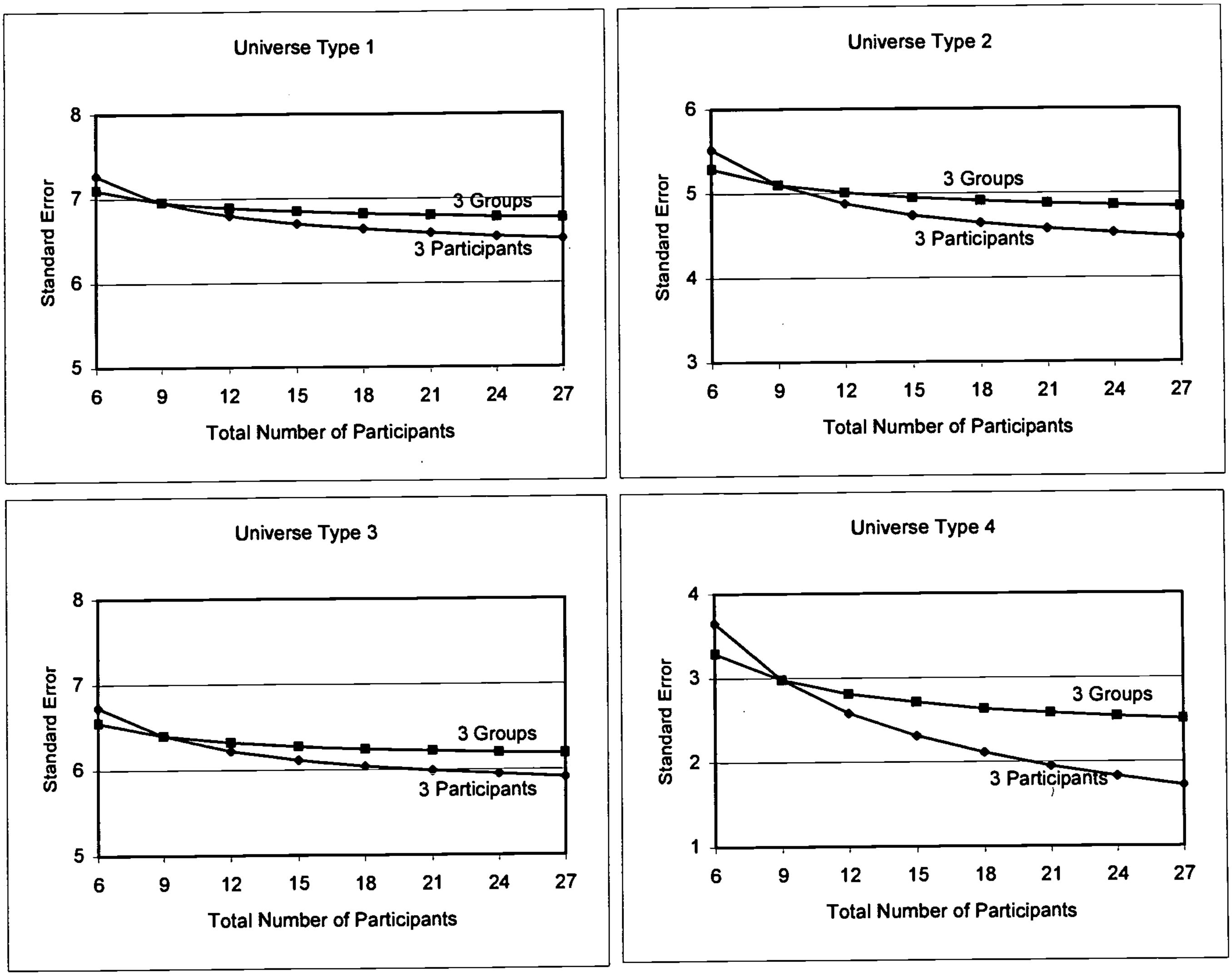


Figure 5. The effects of small groups and participants within small group on standard errors for the Bookmark Cutscores for Given Total Number of Panels (Grade 8).

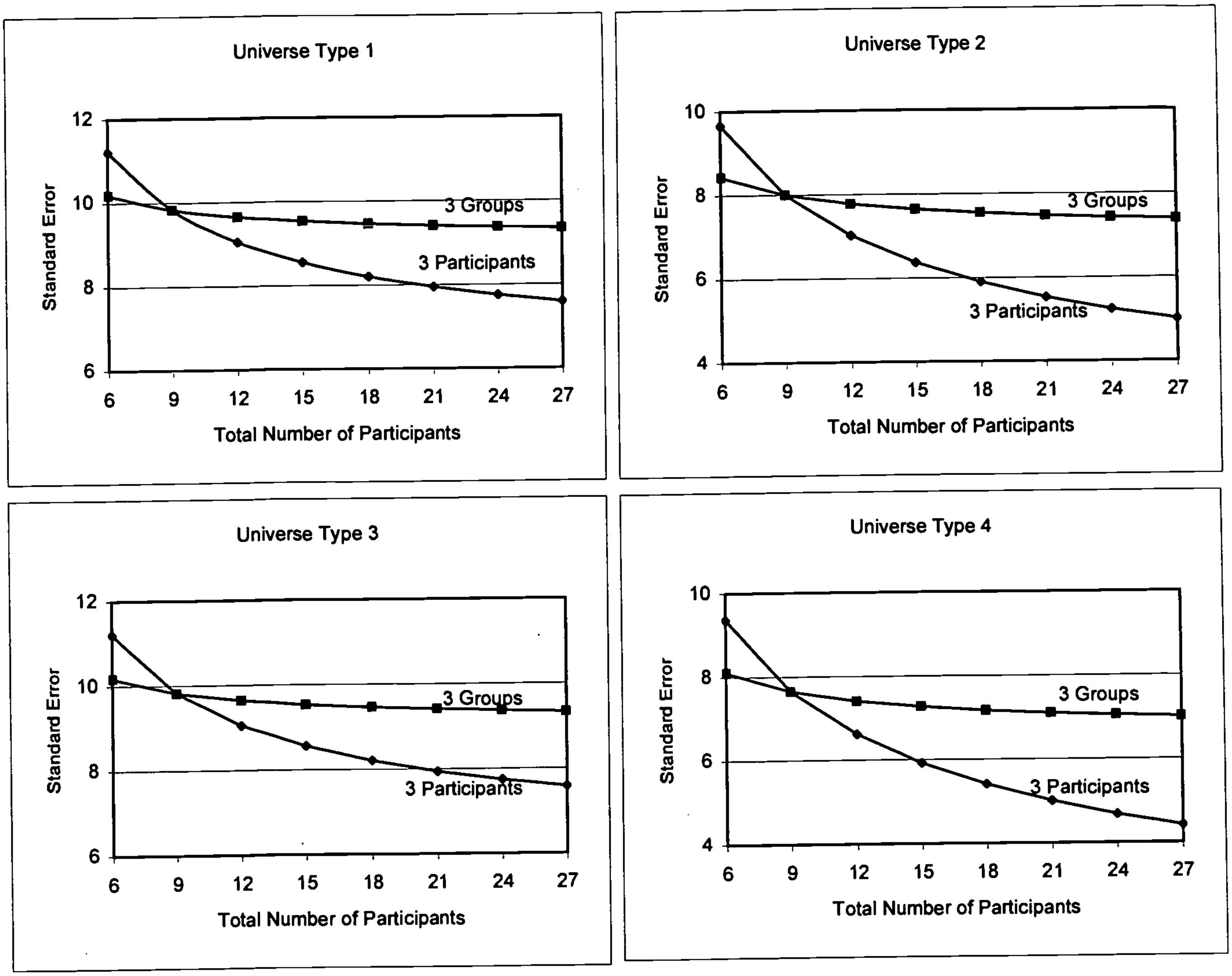
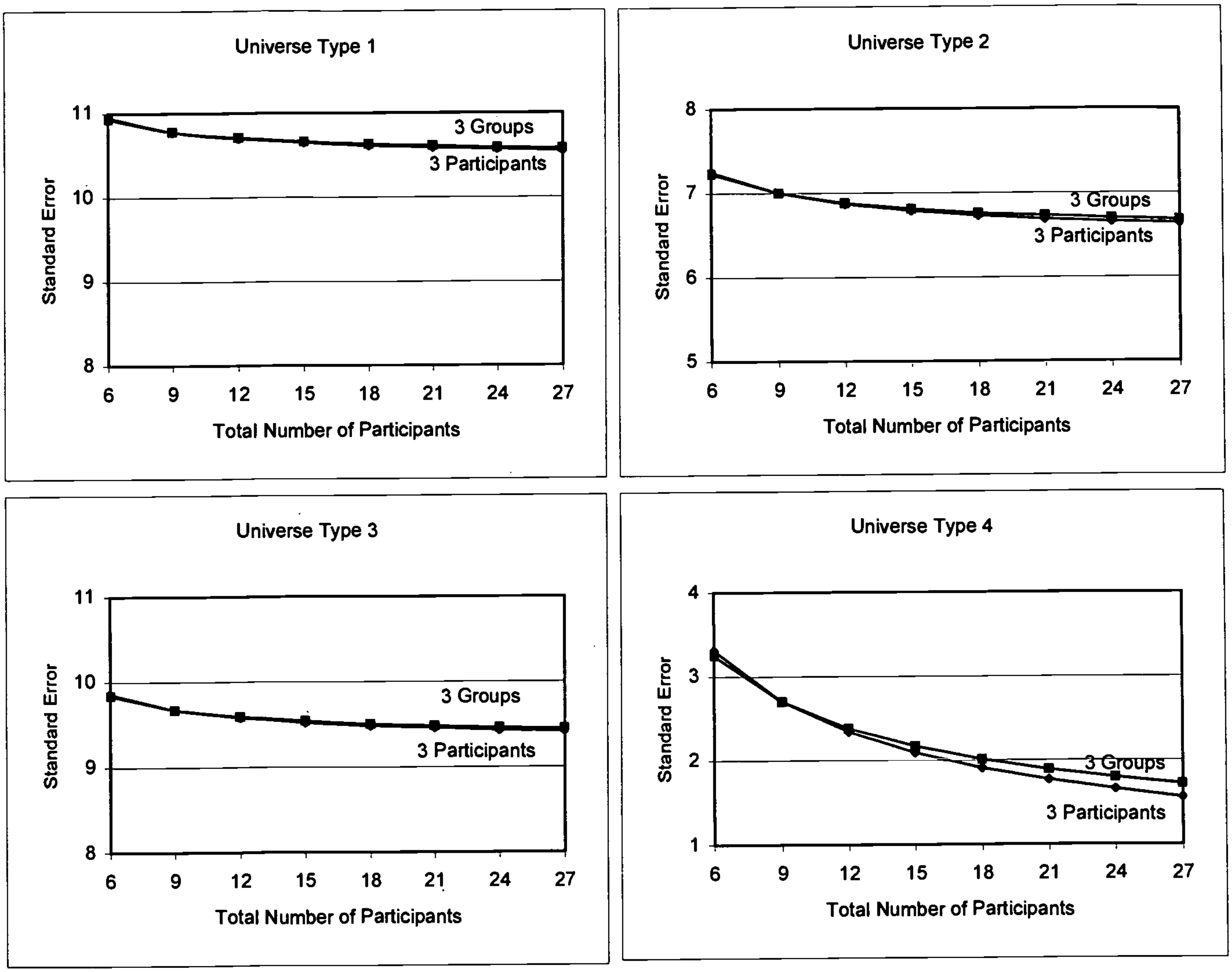


Figure 6. The effects of small groups and participants within small group on standard errors for the Bookmark cutscores for Given Total Number of Panels (Grade 11).


Universe Type 1

Universe Type 2

Universe Type 3

Universe Type 4

Figure 7. The effects of small groups and participants within small group on standard errors for the Bookmark cutscores for Given Total Number of Panels (Grade 8 - Cutscore 3).

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: A generalizability theory approach toward estimating standard errors of cut scores set using the Bookmark standard setting procedure	
Author(s): Guemin Lee and Daniel M. Lewis	
Corporate Source: CTB/McGraw-Hill	Publication Date: April 11, 2001

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	**Level 2A**	**Level 2B**
↑ ☑	↑ □	↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only
Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.		

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: Guemin Lee	Printed Name/Position/Title: Guemin Lee, Research Scientist	
Organization/Address: CTB/McGraw-Hill 20 Ryan Ranch Road Monterey, CA 93940	Telephone: (831) 393-7745	Fax: (831) 393-7016
	E-mail Address: glee@ctb.com	Date: March 13, 2001

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory (Bldg 075)
College Park, Maryland 20742

Telephone: 301-405-7449
Toll Free: 800-464-3742
Fax: 301-405-8134
ericae@ericae.net
http://ericae.net