ED 442 859		TM 031 279
AUTHOR	Lee, Guemin; Sykes, Robert C.	
TITLE	A Comparison of Scoring Modalities for Performance Assessment: A Case of Constructed Response Items.	
PUB DATE	2000-04-24	
NOTE	21p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).	
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)	
EDRS PRICE	MF01/PC01 Plus Postage.	
DESCRIPTORS	Comparative Analysis; *Constructed Response; *Performance Based Assessment; Scores; *Scoring	
IDENTIFIERS	Accuracy	

ABSTRACT

This study was designed to investigate several scoring modalities in terms of the accuracy of scores and the efficiency of scoring procedures. Five scoring modalities (SM) are conceptualized by considering tasks and raters as sources of error where "p," "t," and "r" represent person, task, and rater, respectively. These are: (1) SM1 [p x t x r]; (2) SM2 [p x (t:r)]; (3) SM3 [p x (r:t)]; (4) SM4 [(p:r) x t]; and (5) SM5 [r:(p x t)]. SMI and SM5 seem to be somewhat unrealistic for practical use in large-scale performance assessments. SM2 would be least preferable among the five scoring modalities conceptualized in this study for large-scale performance assessments. SM3 could be the best option when a high level of score accuracy is required and a high level of rater variation might be expected. If considering both score accuracy and efficiency of scoring procedures, SM4 should be considered for use in large-scale performance assessments. (Contains 2 figures, 3 tables, and 10 references.) (Author/SLD)

A Comparison of Scoring Modalities for Performance Assessment: A Case of Constructed Response Items

Guemin Lee
Robert C. Sykes
CTB/McGraw-Hill

Paper Presented at the Annual Meeting
of the American Educational Research Association
New Orleans, LA
April 24, 2000

Abstract

This study was designed to investigate several scoring modalities in terms of (a) the accuracy of scores and (b) the efficiency of scoring procedures. Five scoring modalities (SM) are conceptualized by considering tasks and raters as sources of error: SM1 [$p \times t \times r$], SM2 [$p \times (t:r)$], SM3 [$p \times (r:t)$], SM4 [$(p:r) \times t$], and SM5 [$r:(p \times t)$], where *p, t,* and *r* represent person, task, and rater, respectively. SM1 and SM5 seem to be somewhat unrealistic for practical use in large-scale performance assessments. SM2 would be least preferable among five scoring modalities conceptualized in this study for large-scale performance assessments. SM3 could be the best option when a high level of score accuracy is required and high level of rater variation might be expected. If considering both score accuracy and efficiency of scoring procedures, SM4 should be considered for use in large-scale performance assessments.

A Comparison of Scoring Modalities for Performance Assessments: A Case of Constructed Response Items

Over the past several years, the use of performance assessments as alternatives or supplements to traditional multiple choice tests has been one of the central features of current testing practices (Shavelson, Baxter, & Gao, 1993; Gao, Shavelson, & Baxter, 1994; Cronbach, Linn, Brennan, & Haertel, 1997). As one type of performance assessment, constructed response (CR) items have been broadly used in large scale testing programs both with and without multiple choice items. The use of CR items should require hand-scoring procedures that can be successfully implemented by raters. Consequently, rater variation among total score variation has been focused by several previous studies that investigated psychometric issues of performance assessments (Linn & Burton, 1994; Brennan & Johnson, 1995; Shevelson, Baxter, & Gao, 1993; Gao, Brennan, & Shavelson, 1994; Gao, Shavelson, & Baxter, 1994; Shavelson, Ruiz-Primo, & Wiley, 1999; Clauser, Clyman, Swanson, 1999).

Sykes, Heidorn, and Lee (1999) demonstrated that the allocation of readers to responses in tasks containing more than one CR item can affect the total constructed response scores. Thus, the allocation of raters to items (or items to raters) is an important issue. However, there have been few studies evaluating whether scores can be produced more efficiently under some scoring modalities than others or whether these scores differ in their accuracy. Consequently, it is not clear which scoring modality is relatively accurate and efficient in scoring student's performance. This study was designed for addressing this issue by comparing several scoring modalities in terms of (a) the accuracy of scores and (b) the efficiency (or cost-effectiveness) of scoring procedures. If different degrees of score accuracy are established across different scoring designs, the more efficient (or cost-effective) scoring design can be considered and administered in large scale performance assessments.

The objectives of this study were to:

1. Conceptualize possible scoring modalities for performance assessment incorporating tasks and raters as sources of error.
2. Estimate variance components for each scoring modality and compute absolute and relative generalizability coefficients and standard errors of measurement.
3. Compute rating case-counts and the frequencies of paper movements between raters for each scoring modality.
4. Investigate the relative appropriateness of each scoring modality in terms of the accuracy of scores and the efficiency of scoring procedures.

Scoring Modality

Five scoring modalities are conceptualized in this study when incorporating tasks and raters as sources of error.

Scoring Modality 1 (SM1)

Under this scoring modality, a rater reads every response to every task of every student. The univariate $p \times t \times r$ design, persons (*p*) crossed with tasks (*t*) and raters (*r*), is appropriate. The linear model for the response of a person to a task read by a rater treats persons as objects of measurement and tasks and raters as random facets. This linear model can be represented as:

$$X_{ptr} = \mu + \mu_p \sim + \mu_t \sim + \mu_r \sim + \mu_{pt} \sim + \mu_{pr} \sim + \mu_{tr} \sim + \mu_{ptr,e} \sim . \qquad (1)$$

The terms on the right-hand side of the equation are the grand mean, person effect, task effect, rater effect, person by task interaction effect, person by rater interaction effect, task by rater interaction effect, and person by task by rater interaction effect confounded with unexplained sources of error, respectively.

Scoring Modality 2 (SM2)

Under this scoring modality, a rater reads each student's responses to a subset of tasks. The univariate $p \times (t : r)$ design, persons (*p*) crossed with tasks (*t*) nested within raters (*r*), is appropriate for this scoring design. The linear model can be represented as:

$$X_{ptr} = \mu + \mu_{p} \sim + \mu_{r} \sim + \mu_{t:r} \sim + \mu_{pr} \sim + \mu_{pt:r,e} \sim . \tag{2}$$

The terms on the right-hand side of the equation are the grand mean, person effect, rater effect, task within rater effect, person by rater interaction effect, and person by task within rater interaction effect confounded with unexplained sources of error, respectively.

Scoring Modality 3 (SM3)

Under this scoring design, several raters read each student's response to the same task, but each task is read by different sets of raters. The univariate $p \times (r : t)$ design, persons (*p*) crossed with raters (*r*) nested within tasks (*t*), is appropriate for this scoring design. The linear model can be represented as:

$$X_{ptr} = \mu + \mu_{p} \sim + \mu_{t} \sim + \mu_{r:t} \sim + \mu_{pt} \sim + \mu_{pr:t,e} \sim . \tag{3}$$

The terms on the right-hand side are the grand mean, person effect, task effect, rater within task effect, person by task interaction effect, and person by rater within task interaction effect confounded with unexplained sources of error, respectively.

Scoring Modality 4 (SM 4)

Under this scoring modality, a rater reads all responses to the tasks of a subset of students (not all students). The univariate $(p : r) \times t$ design, persons (*p*) nested within raters (*r*) crossed with tasks (*t*), is appropriate for this situation. The linear model can be represented as:

$$X_{ptr} = \mu + \mu_{p:r} \sim + \mu_{r} \sim + \mu_{t} \sim + \mu_{rt} \sim + \mu_{pt:r,e} \sim . \tag{4}$$

The terms on the right are the grand mean, person within rater effect, rater effect, task effect, rater by task interaction effect, and person by task within rater interaction effect confounded with unexplained sources of error, respectively.

Scoring Modality 5 (SM5)

Under this scoring modality, a rater reads only one student's response to a task. The univariate $r:(p \times t)$ design, raters (r) nested within persons (p) and tasks (t), is appropriate for this scoring design. This linear model can be represented as:

$$X_{ptr} = \mu + \mu_p \sim + \mu_t \sim + \mu_{pt} \sim + \mu_{r:pt,e} \sim . \qquad (5)$$

The terms on the right-hand side of the equation are the grand mean, person effect, task effect, person by task interaction effect, and rater within person and task effect confounded with unexplained sources of error, respectively.

Method

Data Source

Samples of approximately 2,000 students were obtained for a Mathematics field test form at each of grades 5, 8, and 10 and for a Reading field test form at each of grades 4, 8, and 10 of a large-state assessment. Stratified random sampling procedures were used to ensure that the selected samples accurately represented the school population of the state. The test used in this study was the grade 8 Mathematics field test composed of 52 multiple-choice items and 11 constructed-response items. Two types of constructed response items were administered: two-point short response (SR) and four-point extended response (ER) items. Nine SR and two ER items were used in the grade 8 Mathematics field test (total score points = 26). Only students' response vectors to these 11 CR items were analyzed in this study.

Raters were trained to implement the scoring rubrics and anchor papers. The scoring efforts focused on careful rater training, as well as the use of rater check sets and "read behinds" to be certain all raters maintain the same scoring standard. For estimating variance components for the fully crossed design (SM1: $p \times t \times r$ design), a subset of students' response vectors was extracted. In this subset, there were 19 different rater sets and each of them was composed of 2 raters from a total of 10 raters. Two raters of each set read all responses to tasks of all students assigned to that rater set (ranging from 15 to 46 students).

Analyses

The application program GENOVA (Crick & Brennan, 1983) was used in this study for estimating variance components for SM1 using the $p \times t \times r$ generalizability study design. A total of 19 independent runs were completed with 19 data files, and the variance components were averaged over 19 runs for obtaining more stable estimates. The variance components for other scoring modalities can be estimated by manipulating the variance component estimates of SM1 by following specified equations in Table 1.

Insert Table 1 About Here

Unlike classical test theory, two different types of error variance are differentiated in G-theory, relative error variance and absolute error variance. The distinction between the two error variances is associated with separate types of decisions, relative and absolute decisions (Brennan, 1992; Cronbach, Gleser, Nanda, & Rajaratnam, 1972), which are somewhat analogous to norm-referenced and criterion-referenced interpretations, respectively. The absolute and relative generalizability coefficients and standard errors of measurement for each scoring modality are computed from using variance component estimates and assumed sample sizes for tasks and raters in decision studies.

The total rating case-counts can be computed by $\sum_{p=1}^{P}\sum_{t=1}^{T} NR_{pt}$, where NR_{pt} represents the number of ratings on task t of person p. The amount of paper movement can be computed by counting the number of times students' response papers move between raters. The $(p:r) \times t$ scoring design does not require any paper moving because a rater reads all responses to the tasks of a subset of students. For the other designs, studcnts' papers must be handed to various degrees from rater to rater. The paper moving frequency can be computed by $\sum_{p=1}^{P}(R_p - 1)$, where R_p represents the number of raters who read a (or several) response(s) of a student p. The smaller rating case-counts and paper moving frequencies, the more efficient (or cost-effective) the scoring procedure could be.

Results

The Accuracy of Scores and Scoring Modality

Table 2 shows variance component estimates from a generalizability study (G-study) for each scoring modality and also presents proportions of variance components compared to total score variance.

Insert Table 2 About Here

The variance components in a G-study for SM 1 represent the observed score variance for a single response of an individual person to a single task as it was evaluated by a single rater. The person variance is an estimate of the variance of students' mean scores over tasks and raters (17.9% among total variance). Similarly, task variance represents the variation of task mean scores over all persons and raters (31.6%). A quite small amount of rater variance was estimated (0.0%), even though the actual estimate was not zero. With respect to the two-way interaction effects, the magnitude of the person by task variance component suggests that there were considerably different rank orderings of persons across tasks (42.2%). Variance component estimates for the other two two-way interactions including the rater facet, person by rater and task by rater interactions, were very small (0.2% and 0.1%, respectively). The variance component term of person by task by rater is the three-way interaction variance component including variance components due to unexplained sources of error (8.1%). The pattern of the magnitude of variance components (e.g., large variation related to task facet and small variation related to rater facet) seems consistent with previous studies (Linn & Burton, 1994; Brennan & Johnson, 1994). However, the variance component for task was relatively larger than in previous studies. In contrast, the residual variance component (person by task by rater interaction) was relatively smaller than in previous studies.

As Brennan (1992, pp. 30-32) indicated, the fully crossed G-study design is most powerful in that the estimated variance components from it can be used to estimate variance components for any possible designs containing nesting facets. The variance components for SM 1 were actually estimated and variance components for the other scoring modalities were estimated by following the specifications given in Table 1 as described in the previous section.

For estimating relative generalizability coefficients (G-coefficients) and standard errors of measurement (SEMs), several decision studies (D-studies) were conducted under some restrictions. One of these restrictions was that the number of tasks was fixed to 12 as was in actual testing. Because the main purpose of the current study was to investigate the several scoring modalities, the task effect, which can be captured by manipulating the number of tasks, was not examined in this study. With the fixed number of tasks, the number of raters in the D-studies varied from one to four. Associated relative G-coefficients and SEMs are presented in Figure 1 using the number of raters in a D-study as the horizontal axis.

Insert Figure 1 About Here

SM 3 and 5 provided higher relative G-coefficients than did other scoring modalities except when $n'_r=1$ where n'_r represents the number of raters in a D-study. In the case of $n'_r=1$, SM4 produced a G-coefficient similar to those for SM 3 and 5. The relative G-coefficients for SM1 were higher than those for the SM 2 and 4 except when $n'_r=1$, where the G-coefficient for SM4 was higher than that for the SM1. Even though SM4 produced the same G-coefficients regardless of sample size of raters in the various D-studies, it provided more accurate scores than SM2. SM 1, 2, 3 and 5 gained in accuracy of scores to some degree by increasing the sample size of raters in the D-studies. However, the increase for SM2 was much lower than that for SM 1, 3 and 5. The same interpretation can be applied to the relative SEMs in the bottom graph of Figure 1, where the horizontal axis represents the number of raters in the D-studies.

Similar trends were found in the analyses of absolute G-coefficients and SEMs that are presented in Figure 2.

Insert Figure 2 About Here

The Efficiency of Scoring Procedures and Scoring Modality

Several indices that can be used to evaluate the efficiency of scoring procedures are presented in Table 3. To produce these indices, the total number of tasks was set to 12 (as was done in the analyses of score accuracy) and the number of students to 2,000.

Insert Table 3 About Here

In SM1, when the number of raters was set to one only one rater read all responses of 2,000 students to the 12 tasks. Consequently, the rater produced a total of 24,000 ratings. Each response to a task taken by a student was read once. Because only one rater read all responses of all students, there was no need to move students' response papers from rater to rater. If considering two raters for this scoring modality, each rater will read all responses of 2,000 students to 12 tasks. That is, each response of a student will be read twice by two different raters. The total ratings increases to 48,000, and 2,000 students' response papers should move from one rater to another rater. The same logic can be applied to the interpretation of the cases of n'_r=3 and n'_r=4.

SM2 provided the same efficiency indices as did SM1 when n'_r=1. However, when the number of raters increased from one to two, the ratings per rater were reduced from 24,000 to 12,000. That is, each of two raters read a half set of the total tasks for all 2,000 students. Consequently, one rater produced just 12,000 ratings. Each task of a student was read once by either of two raters. For example, one rater read tasks 1-6, and another rater read tasks 7-12. Even though the total ratings did not change regardless of the number of raters, the students' response papers would have to be moved among raters.

In SM3, relatively many raters were required for evaluating students' responses because raters were nested within tasks. (SM5 needs more raters than does SM3.) Setting the number of raters to one does not mean that one rater read all of the students' responses. It rather meant that one task was read by one rater and that different tasks were read by different raters. Consequently, this scoring design also required a large amount of movement of students' paper among raters. For example, in the case of $n'_r=1$, the 2,000 students' response papers has to be passed among 12 raters.

In the case of $n'_r=1$, SM4 produced the same efficiency measures as did SM 1 and 2. When n'_r increased from one to two, one rater read all responses of 1,000 students to the 12 tasks and another rater read all responses of the remaining 1,000 students to the 12 tasks. Thus, this scoring modality did not require students' paper movement among raters. The total ratings remained the same regardless of the number of raters. However, the ratings per rater decreased as the number of raters increased.

In SM5, each rater read one response to one task of a student. Thus, so many raters [(the number of tasks) x (the number of students)] were needed in the case of $n'_r=1$. When using $n'_r=2$, two times more raters were needed. This scoring design also required as much movement of student papers among raters as did SM3. The total ratings increased as the number of raters increased. The number of ratings per rater remained just one regardless of the number of raters.

Discussion

If different degrees of score accuracy were established across different scoring modalities, the more efficient scoring modality can be administered in scoring large scale performance assessment. Among five scoring modalities conceptualized in this study, SM1, 2, 3, and 5 produced more accurate scores by adding more raters. However, the incremental increase for SM2 was much lower than that for SM 1,3, and 5. The score accuracy of SM4 does not depend upon the number of raters. Even though the accuracy of scores in SM2 can be improved by increasing the number of raters, SM4 produces more accurate scores than SM2.

In term of the efficiency of scoring procedures, SM4 seems the most efficient scoring design for evaluating students' performance. SM3 and 5 need more raters and fairly large amounts of movement of student papers among raters. The number of total ratings increases in the SM 1, 3, and 5 as the number of raters increases. This fact can explain the relative high level of score accuracy for SM 1, 3, and 5 compared to SM2 and 4. That is, for these scoring modalities, adding raters results in more ratings per task (more total ratings) and consequently, a high level of score accuracy.

Issues about the relative appropriateness of each scoring modality for specific measurement procedures should be worthwhile to be addressed. In terms of the number of total raters needed, SM5 does not seem to be appropriate and realistic for large-scale performance assessments. This scoring modality may be considered when administering extremely small numbers of tasks and examinees. Even though these conditions are met, the accuracy of scores is almost the same as that from SM3 that needs fewer raters.

For the situation that requires high level of score accuracy for the performance assessments, SM3 can be considered. This scoring design needs more raters than SM 1, 2, and 4, but fewer raters than SM5. The ratings per rater are relatively few compared to other scoring modalities. One problem of this scoring design is associated with the movement of student response papers among raters. SM3 requires such a large amount of paper movement that it makes hard to consider this scoring modality for large-scale performance assessments. However, when a high level of score accuracy is needed and a high level of rater variation is expected, this scoring modality would be worth of consideration.

SM1 produces less accurate scores, but is more efficient than SM3 and 5. The main drawback of this scoring modality is related to the number of ratings per rater. Because every rater should read all the responses of all students to all tasks, this scoring modality puts a great burden on each of the raters. Consequently, this design has not been frequently considered for large-scale performance assessments that involve a large number of examinees and tasks. This design may be considered for a reasonable size of students and a small number of tasks.

In term of score accuracy, SM2 would be least preferable. From the perspective of the efficiency of scoring procedures, this modality does not have any great advantages compared to the other scoring modalities either. Even though this scoring modality lacks advantages of accuracy or efficiency, it has been considered and used in large-scale performance assessments.

SM4 has probably received the lease attention in previous studies dealing with issues associated with scoring modalities. In terms of the efficiency of scoring procedures, this scoring design seems most efficient and cost-effective. From the perspective of score accuracy, this modality is better than SM2 and is not much worse than SM 1, 3, and 5. If considering both score accuracy and efficiency of scoring procedures, this scoring design should be considered for large-scale performance assessments.

The implications discussed above apply in other situations where high generalizability across raters is present. If decisions about scoring modalities have to be made under the situation of high level of rater variation, the above implications should be cautiously considered or different information should be gathered for decision making. Also, it should be mentioned that the Mathematics test was used in this study in estimating variance components. The Mathematics test could be considered one of subject areas that contain the least amount of halo effects (consequently, less variation estimates related to rater facet).

References

Brennan, R.L., & Johnson, E.G. (1995). Generalizability of performance assessments. Educational Measurement: Issues and Practice, 14, 9-12.

Clauser, B.E., Clyman, S.G., & Swanson, D.B. (1999). Components of rater error in a complex performance assessment. Journal of Educational Measurement, 36, 29-45.

Crick, J.E., & Brennan, R.L. (1983). Manual for GENOVA: A generalized analysis of variance system (ACT Technical Bulletin No. 43). Iowa City, IA: ACT.

Cronbach, L.J., Linn, R.L., Brennan, R.L., & Haertel, E. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. Educational and Psychological Measurement, 57, 373-399.

Gao, X., Brennan, R.L., & Shavelson, R.J. (1994, April). Generalizability of group means for performance assessments under a matrix sampling design. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Gao, X., Shavelson, R.J., Baxter, G.P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. Applied Measurement in Education, 7, 323-342.

Linn, R.L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. Educational Measurement: Issues and Practice, 13, 5-8.

Shavelson, R.J., Baxter, G.P., & Gao, X. (1993). Sampling variability of performance assessments. Journal of Educational Measurement, 30, 215-232.

Shavelson, R.J., Ruiz-Primo, M.A., & Wiley, E.W. (1999). Note on sources of sampling variability in science performance assessments. Journal of Educational Measurement, 36, 61-71.

Sykes, R.C., Heidorn, M., & Lee, G. (1999, April). The assignment of raters to items: Controlling for rater effects. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

Table 1
Estimating Variance Components of Scoring Modalities 2 to 5 Using Variance Components Estimates of Scoring Modality 1

SM 1 $p \times t \times r$	SM 2 $p \times (t:r)$	SM 3 $p \times (r:t)$	SM 4 $(p:r) \times t$	SM 5 $r:(p \times t)$
$\hat{\sigma}^2(p)$	$\hat{\sigma}^2(p)$	$\hat{\sigma}^2(p)$	$\hat{\sigma}^2(p:r) = \hat{\sigma}^2(p) + \hat{\sigma}^2(pr)$	$\hat{\sigma}^2(p)$
$\hat{\sigma}^2(t)$	$\hat{\sigma}^2(t:r) = \hat{\sigma}^2(t) + \hat{\sigma}^2(tr)$	$\hat{\sigma}^2(t)$	$\hat{\sigma}^2(t)$	$\hat{\sigma}^2(t)$
$\hat{\sigma}^2(r)$	$\hat{\sigma}^2(r)$	$\hat{\sigma}^2(r:t) = \hat{\sigma}^2(r) + \hat{\sigma}^2(tr)$	$\hat{\sigma}^2(r)$	
$\hat{\sigma}^2(pt)$		$\hat{\sigma}^2(pt)$		$\hat{\sigma}^2(pt)$
$\hat{\sigma}^2(pr)$	$\hat{\sigma}^2(pr)$			
$\hat{\sigma}^2(tr)$			$\hat{\sigma}^2(tr)$	
$\hat{\sigma}^2(ptr,e)$	$\hat{\sigma}^2(pt:r,e) = \hat{\sigma}^2(pt) + \hat{\sigma}^2(ptr,e)$	$\hat{\sigma}^2(pr:t,e) = \hat{\sigma}^2(pr) + \hat{\sigma}^2(ptr,e)$	$\hat{\sigma}^2(pt:r,e) = \hat{\sigma}^2(pt) + \hat{\sigma}^2(ptr,e)$	$\hat{\sigma}^2(r:pt,e) = \hat{\sigma}^2(r) + \hat{\sigma}^2(pr) + \hat{\sigma}^2(rt) + \hat{\sigma}^2(ptr,e)$

Note: SM = Scoring Modality.

Table 2

Variance Component Estimates for Each Scoring Modality in a Generalizability Study

Variance Component	Estimate	Proportion of Variance Component
	Scoring Modality 1	
person (p)	13.3	17.9
task (t)	23.5	31.6
rater (r)	0.0	0.0
person x task (pt)	31.4	42.2
person x rater (pr)	0.2	0.2
task x rater (tr)	0.1	0.1
person x task x rater (ptr)	6.0	8.1
	Scoring Modality 2	
person (p)	13.3	17.9
task within rater (t:r)	23.5	31.6
rater (r)	0.0	0.0
person x rater (pr)	0.2	0.2
person x task within rater (pt:r)	37.4	50.3
	Scoring Modality 3	
person (p)	13.3	17.9
task (t)	23.5	31.6
rater within task (r:t)	0.1	0.1
person x task (pt)	31.4	42.2
person x rater within task (pr:t)	6.2	8.3
	Scoring Modality 4	
person within rater (p:r)	13.4	18.1
task (t)	23.5	31.6
rater (r)	0.0	0.0
task x rater (tr)	0.1	0.1
person x task within rater (pt:r)	37.4	50.3
	Scoring Modality 5	
person (p)	13.3	17.9
task (t)	23.5	31.6
person x task (pt)	31.4	42.2
rater within person x task (r:pt)	6.2	8.4

Note. The scale of the variance component estimates was changed by multiplying all entries by 100 and then rounding to one decimal place.

Table 3
Efficiency Indices of Scoring Modality in Scoring 12 Tasks and 2,000 Examinees

n_r'	Total Raters	Ratings per Rater	Ratings per Task	Total Ratings	Paper Moving
		Scoring Modality 1			
1	1	24,000	1	24,000	0
2	2	24,000	2	48,000	2,000
3	3	24,000	3	72,000	4,000
4	4	24,000	4	96,000	6,000
		Scoring Modality 2			
1	1	24,000	1	24,000	0
2	2	12,000	1	24,000	2,000
3	3	8,000	1	24,000	4,000
4	4	6,000	1	24,000	6,000
		Scoring Modality 3			
1	12	2,000	1	24,000	22,000
2	24	2,000	2	48,000	46,000
3	36	2,000	3	72,000	70,000
4	48	2,000	4	96,000	94,000
		Scoring Modality 4			
1	1	24,000	1	24,000	0
2	2	12,000	1	24,000	0
3	3	8,000	1	24,000	0
4	4	6,000	1	24,000	0
		Scoring Modality 5			
1	24,000	1	1	24,000	22,000
2	48,000	1	2	48,000	46,000
3	72,000	1	3	72,000	70,000
4	96,000	1	4	96,000	94,000

Note. n_r' is the sample size of rater facet in a decision study.

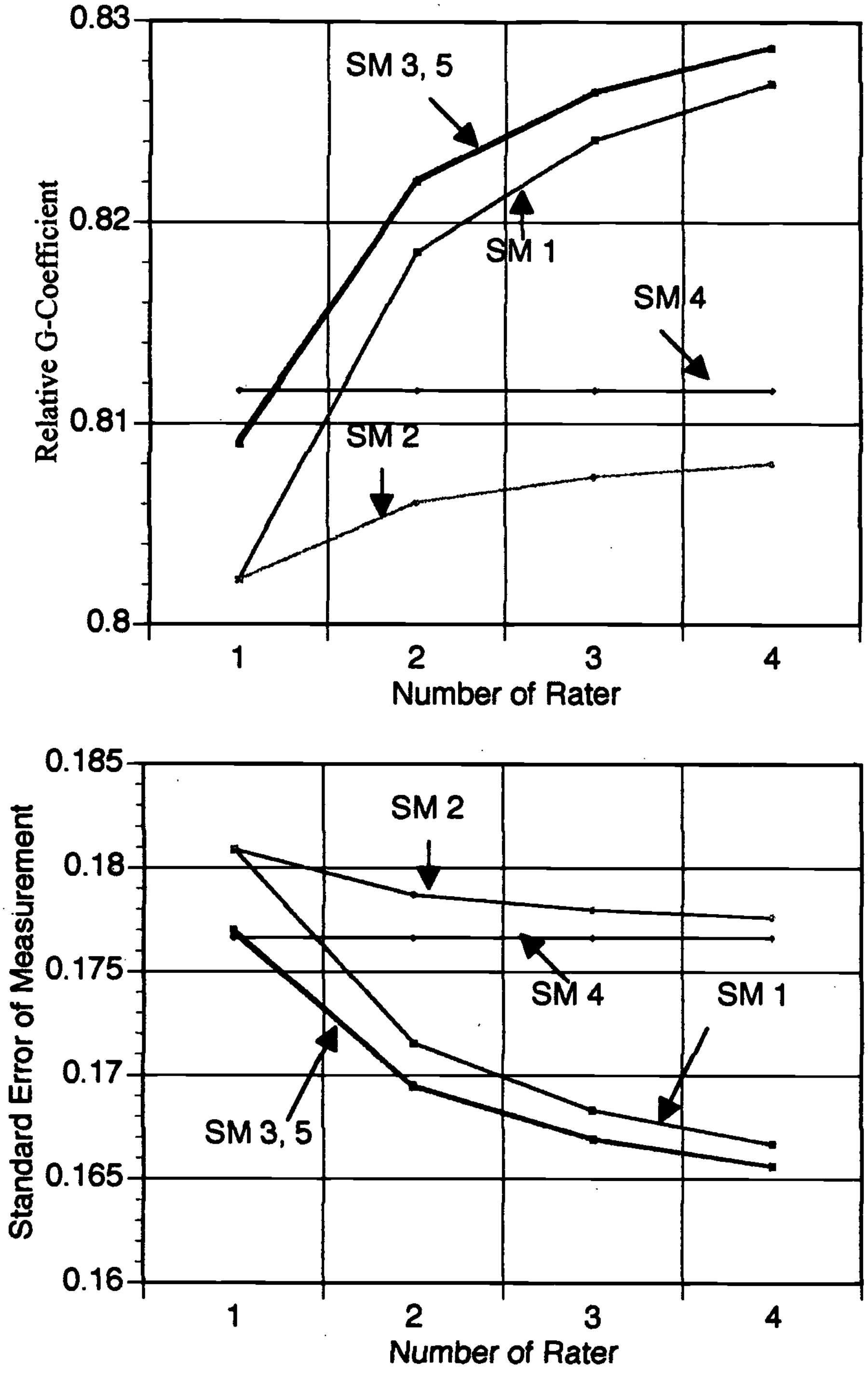

Figure 1. Relative generalizability coefficient and standard error of measurement with fixed number of tasks and varying number of raters

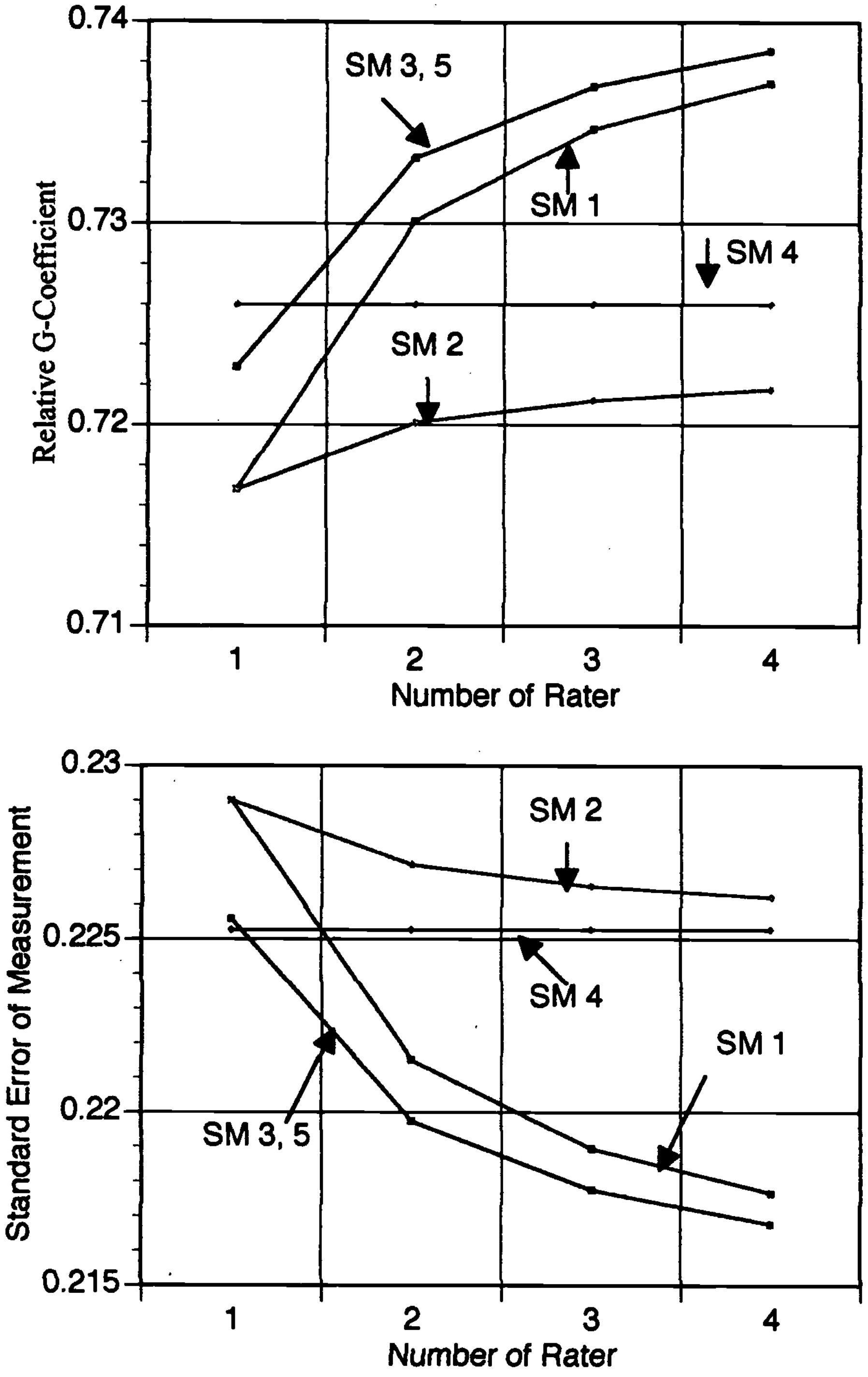

Figure 2. Absolute generalizability coefficient and standard error of measurement with fixed number of tasks and varying number of raters

AERA

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

ERIC®

TM031279

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: A comparison of scoring modalities for performance assessment: A case of constructed response items.

Author(s): Guemin Lee. and Robert C. Sykes

Corporate Source: CTB/McGraw-Hill

Publication Date: April, 2000

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 at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

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)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

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.

Sign here,→ please

Signature: Guemin Lee

Printed Name/Position/Title: Guemin Lee

Organization/Address: CTB/McGraw-Hill
20 Ryan Ranch Road, Monterey CA9394

Telephone: 831-393-7745

FAX: 831-393-7016

E-Mail Address: glee@ctb.com

Date: May 18, 2000

(over)