ACT Research Report Series 96-1

Sampling Variability and Generalizability of Work Keys Listening and Writing Scores

Xiaohong Gao

May 1996

For additional copies write;
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243-0168

Sampling Variability and Generalizability of Work Keys Listening and Writing Scores

Xiaohong Gao

Abstract

The use of the Work Keys Listening and Writing assessment needs to be accompanied by systematic evaluation of its technical qualities. This study examines sampling variability and generalizability of Listening and Writing scores when multiple forms, raters, and prompts (tasks) are used. Different types of scoring, current level scores and mean scores (unrounded and rounded), were also compared in terms of generalizability.

The results indicate that (a) examinees' scores vary from one test form to another due to large task-sampling variability, (b) the rank orderings of prompt difficulty differ for the various examinees, (c) measurement errors are mainly introduced by task-sampling variability not by rater-sampling variability, (d) the Writing scores are more generalizable than the Listening scores, and (e) current level scores are less generalizable than mean scores (both unrounded and rounded). The use of six prompts and two raters in the Work Keys assessment leads to greater generalizability, especially in Writing, than other performance assessments with less numbers of tasks and/or raters. In addition, these analyses also led Work Keys to explore alternative level scores with higher reliability than the current level scores.

Acknowledgments

The author would like to thank Mike Kolen, Dean Colton, and Joy McLarty for their helpful comments about this report and Robert Brennan for his suggestions about generalizability analyses.

Sampling Variability and Generalizability of Work Keys Listening and Writing Scores

The Work Keys assessment system serves as a yardstick for measuring individuals' generic employability skills that are considered critical to success in a wide range of occupations. The current system consists of eight operational assessments including a performance-based assessment: Listening (L) and Writing (W).

The use of the Work Keys Listening and Writing assessment needs to be accompanied by systematic evaluation of its technical qualities. Research on the sampling variability and generalizability of performance assessments has indicated that (a) an individual's performance score varies greatly from one task to another, (b) a large number of tasks are needed to obtain a generalizable measure of an individual's performance, and (c) well-trained raters can provide reliable ratings (Brennan, Gao, & Colton, 1995; Gao, Shavelson, & Baxter, 1994; Shavelson, Baxter, & Gao, 1993). However, in most performance assessments, individuals take only one test form. It is not clear whether or not the individuals performance scores are consistent from one test form to another. In other words, are test forms designed to measure the same construct interchangeable? If people are willing to make decisions based on a single-form score, it is essential to investigate form-sampling variability. Furthermore, when multiple test forms, raters, and tasks are used, as is the case with the Work Keys assessment, it is important to examine the magnitude of sampling variability associated with those sources and their impact on measurement errors and generalizability.

This report presents results from generalizability analyses that examine sampling variability and generalizability of the Work Keys Listening and Writing measurement procedures used in a 1993 pilot study. It provides information about the sampling variability of forms, raters, and prompts (tasks), as well as the estimated effects of using different numbers of forms, raters, and tasks on measurement errors and score generalizability.

More specifically, the study addresses the following questions: (a) What are the major sources of measurement errors associated with the Work Keys measurement procedures: forms, raters, and/or prompts (tasks)? (b) How much error could be expected if the measurement procedures were changed in various ways (e.g., using different numbers of raters, tasks, and/or forms)? (c) What are the effects of changing measurement procedures on score generalizability? (d) How does the generalizability of current level scores compare to that of mean scores (unrounded and rounded)? These questions are addressed within the framework of univariate generalizability (G) theory in which each of the two scores (Listening and Writing) is considered separately.

Method

Data

There were two samples of examinees in the 1993 pilot study. Two forms of the Work Keys Listening and Writing assessment were administered to each examinee. One sample contained 167 examinees (Sample I) who took Form 10cc and Form 11cc, and another sample had 89 examinees (Sample II) who took Form 10cc and Form 12cc. Although a plan was made to randomly assign examinees to the two samples (I and II) and to counter-balance the two test forms within each sample, the procedure was not followed strictly during the test administration. For a given form and sample, three raters assigned Listening scores to all six prompts for all examinees. For the same form and sample, a different group of three raters assigned Writing scores to all six prompts for the examinees. The groups of Listening raters and Writing raters were different for each form and each sample.

Instrument

An important feature of the Work Keys Listening and Writing assessment is that it is carried out in a single test administration but provides two different performance scores--Listening and Writing. The Listening score indicates an examinee's skill at listening to and understanding work-related messages, whereas the Writing score indicates the examinee's skill at writing work related messages. The assessment is administered via an audio tape that contains all directions and messages (prompts). Examinees are asked to listen to six audio-taped prompts ranging from shorter and easier to longer and more complex. After listening to each recorded prompt, examinees are told to construct a written summary of the prompt. The written responses are scored separately for listening and writing skills. The listening score is based on the accuracy and completeness of the information in the examinee's written responses, and the writing score is based on the writing mechanics (such as sentence structure and grammar) and writing style used in the examinee's written responses. All scoring is done by three raters in the situation reported here, but by two raters in operational scoring. The scores range from 0 to 5 for each prompt. Operationally, to earn an overall score at a particular level (i.e., the current level score) for Listening or Writing, the examinee must respond to all of the prompts and at least 9 of the 12 scores (6 prompts x 2 raters) assigned by the raters must be at or higher than that level score.

Design and Analysis

Three designs were used to conduct generalizability analyses of the Work Keys Listening and Writing assessment: person x (rater x task):form, person x rater x task, and person x form. Different types of scoring--current level scores and mean scores (unrounded and rounded)--were compared. The analyses examined sampling variability, standard error of measurement and score generalizability. G theory is treated extensively by Cronbach, Gleser, Nanda, and Rajaratnam (1972), Brennan (1992), and Shavelson and Webb (1991). No attempt is made here to explain generalizability theory in detail, and readers are referred to these references for such explanations.

Estimating variance components. An important contribution of generalizability (G) theory to measurement theory is that it allows people to disentangle multiple sources of measurement error. Using analysis of variance (ANOVA) methods, the magnitudes of the sources of variability can be estimated as variance components. In presenting their theory of generalizability, Cronbach et al. (1972) introduced distinctions between what they called generalizability (G) studies and decision (D) studies. "A G study collects data from which estimates can be made of the components of variance for measurements made by a certain procedure; a D study collects data for the purpose of making decisions or drawing conclusions (Cronbach et al., 1972, p.16)." In most applications, G studies attempt to identify and to estimate as many potential sources of variation as possible; D-study considerations, instead of D studies, use the information obtained by the G studies to estimate variance components associated with specific measurement procedures and to design cost-efficient measures for particular purposes. In the present report, various D-study considerations are discussed.

In generalizability analyses, the best estimate of measurement error and generalizability is one that reflects the impact of all sources of variability (Feldt & Brennan, 1989). The purpose of the Work Keys pilot study was to examine sources of variability related to forms, raters, and prompts (tasks). The original data collection design contained three facets--forms, raters, and tasks. More specifically, the examinees (p) took two test forms (f), each form contained six prompts or tasks (t), and each written response was scored by three raters (r). Therefore, a complete G-study design for Listening or Writing was person x [(rater x task):form]. The symbol "x" designates a crossed effect and the ":" means a nested effect. All of these factors are considered to be random effects.

For the first and complete G-study design in Listening and Writing, the raw score assigned to any person (p) by any rater (r) based on any tasks (t) within any form (f) can be presented as:

$$X_{(prt:f)} = \mu + \mu_{p}{\sim} + \mu_{f}{\sim} + \mu_{r:f}{\sim} + \mu_{t:f}{\sim} + \mu_{pf}{\sim} + \mu_{pr:f}{\sim} + \mu_{pt:f}{\sim} + \mu_{rt:f}{\sim} + \mu_{prt:f}{\sim}. \quad (1)$$

The term μ is the grand mean (or expected value) over all persons in the population and all forms, raters, and tasks in the "universe of admissible observations;" the number of persons in the population is infinite, and the numbers of forms, raters, and tasks in the universe of admissible observations are infinite. The other nine terms in Equation 1 are called "score effects." They are defined in terms of mean scores. For example, the person score effect is

$$\mu_{p}{\sim} = \mu_p - \mu,$$

where μ_p is the mean (or expected value) for a person over all forms, raters, and tasks in the universe of admissible observations.

The total variance of the observed scores in Equation 1 is

$$\sigma^2(X_{prt:f}) = \sigma^2_p + \sigma^2_f + \sigma^2_{r:f} + \sigma^2_{t:f} + \sigma^2_{pf} + \sigma^2_{pr:f} + \sigma^2_{pt:f} + \sigma^2_{rt:f} + \sigma^2_{prt:f,e} \qquad (2)$$

where the nine terms to the right of the equal sign are called "variance components." The notation for the residual variance component ($\sigma^2_{prt:f,e}$) reflects the confounding of p x r x t interaction with other unidentified sources of variation (here used *e*). The variance components provide a decomposition of $\sigma^2(X_{prt:f})$ and allow us to examine the magnitudes of sampling variabilities of forms, raters, and tasks, as well as the interactions.

It is important to note that $\sigma^2(X_{prt:f})$ is not the variance for person total scores or mean scores over the 2 forms, 3 raters, and 6 tasks used in the data collection. Rather, $\sigma^2(X_{prt:f})$ is the variance for the scores obtained by single persons on single tasks scored by single raters within single forms. Consequently, the variance components in Equation 2 are also for single person-form-rater-task scores. For example, $\sigma^2_{pt:f}$ is to be interpreted as the variance of the $\mu_{pt:f}\sim$ effect for single person-task interactions within forms. The magnitudes of sampling variability, averaging over numbers of conditions (e.g., number of tasks), are estimated within the contexts of measurement procedures (or universes of generalization) in decision studies or D-study considerations.

The second G-study design, person x rater x task, is applied to each form of the Listening and Writing assessment. The raw score assigned to any person (p) by any rater (r) based on any prompt or task (t) can be represented as:

$$X_{prt} = \mu + \mu_p\sim + \mu_r\sim + \mu_t\sim + \mu_{pr}\sim + \mu_{pt}\sim + \mu_{rt}\sim + \mu_{prt}\sim \qquad (3)$$

and the total variance of the observed score in Equation 3 is

$$\sigma^2(X_{prt}) = \sigma^2_p + \sigma^2_r + \sigma^2_t + \sigma^2_{pr} + \sigma^2_{pt} + \sigma^2_{rt} + \sigma^2_{prt,e}\,. \qquad (4)$$

This design allows the examination of the variances of raters, tasks, and interactions for the scores obtained by single persons on single tasks scored by single raters. It is worth noting that form is a hidden facet in this design.

Furthermore, Work Keys reports the Listening and Writing level scores to individuals and to educational and business agencies. In the present study, each examinee took two forms and obtained two level scores. These level scores are aggregated over raters and tasks. Thus, for the third G-study design (i.e., person x form) the level score or mean score (unrounded or rounded) assigned to any person (p) based on any form (f) can be represented as:

$$X_{pf} = \mu + \mu_{p}{\sim} + \mu_{f}{\sim} + \mu_{pf}{\sim} \qquad (5)$$

and the total variance of the observed score in Equation 5 is

$$\sigma^2(X_{pf}) = \sigma^2_{p} + \sigma^2_{f} + \sigma^2_{pf,e}. \qquad (6)$$

This type of G study allows us to look at form-sampling variability when level scores or mean scores (unrounded or rounded) are used.

The variance components in Equations 2, 4, and 6 can be estimated using analysis of variance (ANOVA) procedures (see Brennan, 1992). A computer program, GENeralized Analysis of VAriance System (GENOVA) developed by Crick and Brennan (1983), was used for obtaining estimated variance components.

Up to this point, all variance components are estimated in G studies and are for single scores (e.g., a single person responding to one random task, scored by one random rater). In practice, decisions about examinees are typically based on total or average scores over some numbers of conditions of the facets (e.g., numbers of raters and tasks). Therefore, it is important to estimate the magnitudes of sampling variabilities associated with a specific measurement procedure in a universe of generalization involving samples of tasks and raters. In generalizability theory, such estimation is carried out in a decision (D) study or a D-study consideration.

For the first design considered here, the average score for an examinee over n'_{f} forms, n'_{r} raters, and n'_{t} tasks can be represented as

$$X_{(pRT:F)} = \mu + \mu_{p}{\sim} + \mu_{F}{\sim} + \mu_{R:F}{\sim} + \mu_{T:F}{\sim} + \mu_{pF}{\sim} + \mu_{pR:F}{\sim} + \mu_{pT:F}{\sim}$$
$$+ \mu_{RT:F}{\sim} + \mu_{pRT:F}{\sim}. \qquad (7)$$

Equation 7 is analogous to Equation 1 with F, R, and T in Equation 7 replacing f, r, and t in Equation 1. Upper-case letters are used to represent means taken over the conditions sampled from the facets in the universe of generalization. Also, the primes in n'_{f}, $n'_{t:f}$, and $n'_{r:f}$ indicate that these D-study sample sizes need not be the same as those used in the G study.

Equation 7 is the model for a p x [(R x T):F] D-study design in which any instance of a measurement procedure is obtained theoretically in the following manner: (a) obtain a random sample of n'_f forms which contain a random sample of tasks, $n'_{t:f}$; (b) obtain a random sample of raters for each form, $n'_{r:f}$; (c) administer all n'_f forms to all persons, p; and (d) have all raters score the responses of all persons to all tasks within a form. This is a verbal description of the three-facet (forms, raters, and tasks) nested design associated with the model in Equation 7. The universe of generalization consists theoretically of all possible instances of the measurement procedure. The instances are considered as "randomly parallel" when they are conceptualized in this manner.

The total variance of the $X_{p(RT:F)}$ scores given by Equation 7 is

$$\sigma^2(X_{pRT:F}) = \sigma^2_p + \sigma^2_F + \sigma^2_{R:F} + \sigma^2_{T:F} + \sigma^2_{pF} + \sigma^2_{pR:F} + \sigma^2_{pT:F} + \sigma^2_{RT:F} + \sigma^2_{pRT:F,e} \quad (8)$$

where the variances to the right of the equal sign are called "D-study variance components." The D-study variance component σ^2_p is called "universe score variance." In this case it is identical to the G-study σ^2_p in Equation 2, and it is analogous to true score variance in classical test theory (CTT). Universe score variance can be conceptualized in the following manner: (a) obtain the universe score for each examinee in the population, where universe score is defined as the mean (or expected value) of observed scores $X_{(pRT:F)}$ over all instances of the measurement procedure in the universe of generalization; and (b) obtain the variance over examinees of these universe scores.

The remaining D-study variance components in Equation 8 are related to the G-study variance components in Equation 2 in the following manner

$$\begin{aligned} &\sigma^2_F = \sigma^2_f \,/\, n'_f, && \sigma^2_{R:F} = \sigma^2_{r:f} \,/\, (n'_{r:f} n'_f), \\ &\sigma^2_{T:F} = \sigma^2_{t:f} \,/\, (n'_{t:f} n'_f), && \sigma^2_{pF} = \sigma^2_{pf} \,/\, n'_f, \\ &\sigma^2_{pR:F} = \sigma^2_{pr:f} \,/\, (n'_{r:f} n'_f), && \sigma^2_{pT:F} = \sigma^2_{pt:f} \,/\, (n'_{t:f} n'_f), \\ &\sigma^2_{RT:F} = \sigma^2_{rt:f} \,/\, (n'_{r:f} n'_{t:f} n'_f), \quad \text{and} && \sigma^2_{pRT:F,e} = \sigma^2_{prt:f,e} \,/\, (n'_{r:f} n'_{t:f} n'_f). \qquad (9) \end{aligned}$$

The eight equations in Equation Set 9 show how the D-study variance components get reduced as D-study sample sizes (n'_f, $n'_{r:f}$, and $n'_{t:f}$) increase. Each of the eight variance components contributes to error variances in measuring examinees' levels of performance.

Analogously, we can obtain the total variance of the X_{pRT} scores given by Equation 4 and the total variance of the X_{pF} scores given by Equation 6, respectively:

$$\sigma^2(X_{pRT}) = \sigma_p^2 + \sigma_R^2 + \sigma_T^2 + \sigma_{pR}^2 + \sigma_{pT}^2 + \sigma_{RT}^2 + \sigma_{pRT,e}^2 \qquad (10)$$

and

$$\sigma^2(X_{pF}) = \sigma_p^2 + \sigma_F^2 + \sigma_{pF,e}^2. \qquad (11)$$

Estimating measurement error. Since an assessment score reflects only a sample of an examinee's performance, it is always subject to sampling error. A standard error (SE) of measurement indicates what spread of results (i.e., observed score range) would be likely if repeated, randomly parallel assessments could collect many scores for the same examinee. It plays an important role in indexing measurement precision. It can be used to form a range in which the examinee's universe score is most likely to be (Cronbach, Linn, Brennan, & Haertel, 1995). It can provide information on the probability (or the percentage) of misclassification of the examinee(s) so that decision makers can decide whether or not the misclassification rate is acceptable for a specific decision. The standard error can also be used to estimate minimum passing and maximum failing scores given a specified standard of proficiency with a certain level of confidence (Linn & Burton, 1994).

In generalizability theory, how generalizable a measurement procedure is depends, in part, on how the scores will be used in making decisions: to rank order examinees (i.e., relative decisions) or to index examinees' levels of performance (i.e., absolute decisions). Two different types of error variance are associated with these two types of decisions: relative error variance and absolute error variance. In this report, interest focuses on the absolute value of an examinee's score (i.e., absolute decisions), and the appropriate error variance is the absolute error variance. For the p x [(R x T):F] D-study design, the absolute error variance can be calculated as:

$$\begin{aligned}\sigma_\Delta^2 &= \sigma_F^2 + \sigma_{R:F}^2 + \sigma_{T:F}^2 + \sigma_{pF}^2 + \sigma_{pR:F}^2 + \sigma_{pT:F}^2 + \sigma_{RT:F}^2 + \sigma_{pRT:F,e}^2 \\ &= \frac{\sigma_f^2}{n'_f} + \frac{\sigma_{r:f}^2}{n'_{r:f}n'_f} + \frac{\sigma_{t:f}^2}{n'_{t:f}n'_f} + \frac{\sigma_{pf}^2}{n'_f} + \frac{\sigma_{pr:f}^2}{n'_{r:f}n'_f} + \frac{\sigma_{pt:f}^2}{n'_{t:f}n'_f} + \frac{\sigma_{rt:f}^2}{n'_{r:f}n'_{t:f}n'_f} + \frac{\sigma_{prt:f,e}^2}{n'_{r:f}n'_{t:f}n'_f}\end{aligned} . \qquad (12)$$

The absolute error variance, σ_Δ^2, is the variance of the difference between examinee observed and universe scores. The square root of the absolute error variance is the standard error of measurement (σ_Δ) for absolute decisions (absolute error). For the p x R x T D-study design, the absolute error variance is

$$\begin{aligned}\sigma_\Delta^2 &= \sigma_R^2 + \sigma_T^2 + \sigma_{pR}^2 + \sigma_{pT}^2 + \sigma_{RT}^2 + \sigma_{pRT,e}^2 \\ &= \frac{\sigma_r^2}{n_r'} + \frac{\sigma_t^2}{n_t'} + \frac{\sigma_{pr}^2}{n_r'} + \frac{\sigma_{pt}^2}{n_t'} + \frac{\sigma_{rt}^2}{n_r' n_t'} + \frac{\sigma_{prt,e}^2}{n_r' n_t'} .\end{aligned} \tag{13}$$

For the p x F D-study design, the absolute error variance can be defined as

$$\sigma_\Delta^2 = \sigma_F^2 + \sigma_{pF}^2 = \frac{\sigma_f^2}{n_f'} + \frac{\sigma_{pf,e}^2}{n_f'} . \tag{14}$$

This report evaluates how much error of measurement or uncertainty to expect if the design is changed in various ways (e.g., by varying the numbers of tasks and raters). Moreover, bands or intervals containing universe scores (e.g., ± 1.645SE) can be computed for different D-study considerations. Percentages (or probabilities) of misclassification as a consequence of measurement error can be estimated as can the impact of increasing the numbers of instances (e.g., raters and tasks) on the percentages. In addition, minimum passing and maximum failing scores given a standard of proficiency can be computed for certain confidence (e.g., 90%) as a function of numbers of raters and tasks.

Estimating generalizability coefficients. Although G theory focuses on the sources of variability that contribute to measurement errors, it also provides reliability-like coefficients: the generalizability (G) coefficient for a relative decision (i.e., relative G coefficient) and the dependability coefficient for an absolute decision (i.e., absolute G coefficient). While magnitudes of estimated error variances (or measurement errors) depend on the scale of measurement used, the reliability-like coefficients do not. This independence of scaling allows us to compare precision of measurement procedures across different scales. This report focuses on the dependability coefficients or absolute G coefficients (Φ) of Work Keys Listening and Writing assessment in making absolute decisions.

A dependability coefficient (Φ) can be viewed as the ratio of universe score variance to total variance that reflects variability in absolute levels of performance. For the Listening and Writing assessment with persons as the objects of measurement, the dependability coefficient for the three types of D-study designs is defined as:

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\Delta^2} \ . \qquad (15)$$

The variance components entering σ_Δ^2 depend on a specific D-study design used. The dependability coefficients show how accurate the generalization is from an examinee's observed score to his/her universe score. In this report, the effects of using different numbers of instances (e.g., forms, raters, and/or tasks) on score generalizability of Listening and Writing assessment are examined.

Results and Discussion

A series of generalizability (G) analyses were conducted to (a) estimate variance components associated with various sources of sampling variation, (b) assess standard errors of measurement and associated decision consistency of different measurement procedures, and (c) examine the generalizability of the Listening and Writing assessment scores. The results provide information about likely psychometric characteristics of the assessment.

Estimated Variance Components

Variance components estimated in generalizability (G) studies show the magnitudes of sampling variabilities in a universe of admissible observations when single scores are used. Estimated variance components in decision (D) studies or D-study considerations indicate the magnitudes of variabilities when generalizing from average or total scores over n' instances to the universe scores in the universe of generalization.

p x [(r x t):f] design. Table 1 provides the estimated G-study variance components ($\hat{\sigma}^2(\alpha)$) and the percents of total variability (%) for Work Keys Listening and Writing scores. The estimates indicate the magnitudes of sampling variation associated with each source (forms, raters, and tasks) and their relative contributions to measurement errors. The person by task interaction contributes most to measurement errors for both Listening and Writing, indicating that the rank orders of examinees vary from one task to another. The finding of a large person by task interaction is consistent with other reported results on performance assessments (see Brennan et al., 1995; Gao et al., 1994; Shavelson et al., 1993). Moreover, the estimated task variance component is the second largest for Listening, suggesting that the tasks within a form differ in difficulty. For example, the task means for Listening 10cc range from 2.383 to 3.473 in Sample I (see Table 2). The results are consistent with the test descriptions which state that the prompts are ordered from easy to difficult. However, tasks do not differ so greatly in difficulty for Writing. For example, the means for Writing 10cc range from 2.764 to 3.200 in Sample I.

TABLE 1

Variance Component Estimates of p x [(r x t):f] Generalizability Studies

	Forms 10cc and 11cc (Sample I)		Forms 10cc and 12cc (Sample II)	
Source of Variability	$\hat{\sigma}^2(\alpha)$	%	$\hat{\sigma}^2(\alpha)$	%
		Listening		
Person (p)	0.26104	21.04	0.20678	17.71
Form (f)	0.04529	3.65	0.05701	4.88
Rater:Form (r:f)	0.00472	0.38	0.00164	0.14
Task:Form (t:f)	0.26973	21.75	0.24257	20.78
pf	0.01767	1.42	0.00639	0.55
pr:f	0.00755	0.61	0.00323	0.28
pt:f	0.47268	38.11	0.45818	39.24
rt:f	0.00338	0.27	0.00498	0.43
prt:f,e	0.15833	12.76	0.18678	16.00
		Writing		
Person (p)	0.37201	45.83	0.29749	36.64
Form (f)	0.00000	0.00	0.00376	0.46
Rater:Form (r:f)	0.00410	0.51	0.00384	0.47
Task:Form (t:f)	0.01136	1.40	0.03338	4.11
pf	0.01964	2.42	0.03357	4.14
pr:f	0.01908	2.35	0.03056	3.76
pt:f	0.23229	28.61	0.26200	32.27
rt:f	0.00353	0.43	0.00038	0.05
prt:f,e	0.14976	18.45	0.14688	18.09

TABLE 2

Means and Standard Deviations Of Work Keys Listening and Writing Tasks

		Listening		Writing	
Task		10cc	11cc/12cc[a]	10cc	11cc/12cc
		Sample I			
1	Mean	3.28942	4.24551	3.19960	3.08982
	SD	1.04744	0.97575	0.99870	0.88452
2	Mean	2.76846	3.31936	3.03194	2.90419
	SD	0.81708	0.77646	0.75715	0.70007
3	Mean	3.47305	3.13972	2.93812	2.91617
	SD	0.99888	0.87608	0.80129	0.68813
4	Mean	2.35729	3.05389	2.97006	2.86228
	SD	0.85851	0.88367	0.78543	0.75804
5	Mean	2.50100	3.05988	2.96008	2.87824
	SD	0.89463	0.91273	0.89533	0.77387
6	Mean	2.38323	2.53693	2.76447	2.82635
	SD	0.79121	0.90429	0.88443	0.91092
		Sample II			
1	Mean	3.34082	3.06367	3.10861	2.88764
	SD	0.98982	1.13693	0.94117	0.83476
2	Mean	2.64419	3.50187	2.77903	2.40075
	SD	0.72265	0.73876	0.78968	0.61991
3	Mean	3.24719	3.39326	2.85393	2.64045
	SD	0.88267	0.68822	0.68342	0.65948
4	Mean	2.17228	2.80899	2.77154	2.55056
	SD	0.82895	0.78478	0.89836	0.70191
5	Mean	2.10487	2.35206	2.64419	2.65169
	SD	0.85811	0.77583	0.92359	0.73497
6	Mean	2.13858	3.20225	2.44944	2.55805
	SD	0.90706	0.87128	1.02260	0.77162

[a] Sample I took Forms 10cc and 11cc and Sample II took Forms 10cc and 12cc.

Further, the form difficulty, averaging over examinees, raters, and prompts, is different for Listening but not for Writing. For example, the mean is 2.795 for Listening Form 10cc but is 3.226 for Listening Form 11cc in Sample I. The average Writing scores are 2.977 for Form 10cc and 2.913 for Form 11cc, respectively. However, the individual scores vary somewhat from one form to another for both Listening and Writing (i.e., person by form interaction). The results suggest that some score adjustment may be needed so that the Listening and Writing scores obtained from different forms are comparable. Meanwhile, traditional equating methods may not be entirely satisfactory here due to some person by form interactions. Furthermore, because forms are not counter-balanced, form-sampling variability is possibly confounded with order and/or practice effects.

For Writing, the universe score variance is larger than the other estimated variance components and is larger than that for Listening, suggesting that there is considerably more variation among examinees with respect to their levels of proficiency in Writing than in Listening. The finding is consistent across Sample I and Sample II. Similar findings were reported on Work Keys data collected in a previous year (see Brennan et al., 1995).

As seen in Table 1, the rater-sampling variability is small, especially for Listening. The fact that rater variance is small means that raters are about equally stringent on average. The fact that the rater by person interaction is small means that examinees are rank ordered about the same by the various raters. The results, thus, suggest that raters are not nearly as large a contributor to total variance as are prompts. It is possible to use a small number of well-trained raters to score each examinee's responses in future operational forms if the training and scoring procedures continue to be well developed and used. It is noteworthy that the variance component (prt:f,e) for a person by rater by task interaction confounded with other unidentified sources of error is relatively large.

The estimates in Table 1 are for single person-rater-task-form scores only. In practice, decisions about examinees are typically made based on average or total scores over some numbers (n') of tasks, raters and/or forms defined by the universe of generalization. Assuming one form, two raters and six tasks are used in the p x [(R x T):F] D studies, Figure 1 at end of the report provides the estimated variance components for the Listening (L) and Writing (W) assessment. Increasing the number of tasks from one to six dramatically decreases the estimated task variance components and the person by task interactions for both Listening and Writing although tasks still count for a large proportion of the total variability.

p x r x t design. To estimate magnitudes of variance components associated with a measurement procedure that contains only raters and tasks as measurement facets, separate p x r x t G studies were carried out for each form. The results show similar patterns for these forms across Sample I and Sample II. Table 3 provides the estimated G-study variance components ($\hat{\sigma}^2(\alpha)$) for single person-rater-task scores. Variance component estimates for task are larger for Listening than for Writing, indicating that the average scores on Listening vary more from one task to another than those on Writing. The large person by task interactions for both Listening and Writing indicates that the rank ordering of prompt difficulty is substantially different for the various examinees. Moreover, the interaction is larger for Listening than for Writing. Again, the estimated components for rater and related interactions are small for the Listening and Writing assessment (except the residuals). The rater by examinee interaction is smaller for Listening than for Writing.

Table 3 also presents the D-study estimates ($\hat{\sigma}^2(\overline{\alpha})$) for average scores if two raters and six prompts (tasks) were used for each form. Figures 2a and 2b plot these estimated variance components. Still, the person by task interaction is the major source of measurement error as compared to the other components for both Listening (L) and Writing (W).

Moreover, it is important to note that the patterns of the estimated variance components are similar across the G studies with different samples of examinees or forms. However, the magnitudes are different from one G study to another due to sampling errors. For example, the magnitudes of the estimated variance components for Form 10cc are different with different samples of examinees and raters. Although the same examinees took both forms the estimated universe score variances are different from one form to another.

TABLE 3

Variance Component Estimates of p x r x t Generalizability Analyses

	Form 10cc		Form 11cc/12cc[a]	
Source of Variability	$\hat{\sigma}^2(\alpha)$	$\hat{\sigma}^2(\bar{\alpha})$[b]	$\hat{\sigma}^2(\alpha)$	$\hat{\sigma}^2(\bar{\alpha})$
		Listening (Sample I)		
Person (p)	0.27024	0.27024	0.28716	0.28716
Rater (r)	0.00858	0.00429	0.00087	0.00043
Task (t)	0.22675	0.03779	0.31272	0.05212
pr	0.00593	0.00296	0.00918	0.00459
pt	0.50007	0.08334	0.44529	0.07422
rt	0.00140	0.00012	0.00537	0.00045
prt,e	0.14610	0.01218	0.17055	0.01421
		Writing (Sample I)		
Person (p)	0.43020	0.43020	0.35312	0.35312
Rater (r)	0.00107	0.00054	0.00714	0.00357
Task (t)	0.01606	0.00268	0.00665	0.00111
pr	0.01802	0.00901	0.02014	0.01007
pt	0.23949	0.03992	0.22509	0.03751
rt	0.00618	0.00051	0.00088	0.00007
prt,e	0.17965	0.01497	0.11988	0.00999
		Listening (Sample II)		
Person (p)	0.22229	0.22229	0.20405	0.20405
Rater (r)	0.00240	0.00120	0.00088	0.00044
Task (t)	0.31605	0.05267	0.16909	0.02818
pr	0.00447	0.00223	0.00199	0.00100
pt	0.47983	0.07997	0.43653	0.07275
rt	0.00012	0.00001	0.00983	0.00082
prt,e	0.15281	0.01273	0.22076	0.01840
		Writing (Sample II)		
Person (p)	0.42761	0.42761	0.23451	0.23451
Rater (r)	0.00285	0.00143	0.00483	0.00241
Task (t)	0.04408	0.00735	0.02268	0.00378
pr	0.02786	0.01393	0.03325	0.01663
pt	0.28726	0.04788	0.23675	0.03946
rt	0.00040	0.00003	0.00035	0.00003
prt,e	0.16864	0.01405	0.12512	0.01043

[a] Sample I took Forms 10cc and 11cc; Sample II took Forms 10cc and 12cc.

[b] D-study variance component estimates were calculated with $n'_r = 2$ and $n'_t = 6$.

p x f design. The previous analyses were conducted on raw scores of the Listening and Writing assessment. The analyses in this section dealt with level scores as well as mean scores (unrounded and rounded). Recall, operationally, to earn an overall level score, the examinee must respond to all of the prompts and at least 9 of the 12 scores assigned by the raters must be at or higher than that level score.

As indicated in Table 4 below and Figure 3, the form variability is notably large for Listening (L), indicating that the two forms (10cc vs. 11cc in Sample I, or 10cc vs. 12cc in Sample II) are not equivalent in their average difficulty (see also Table 2) and some score adjustment may be needed to generate equivalent scores across the forms. However, the form variance component estimates for Writing (W) are negligible for the two samples. The results are consistent with those reported earlier in the p x [(r x t):f] generalizability analyses with raw scores. Moreover, the large person by form interactions for both Listening and Writing level scores strongly suggest that the individuals' performances on the two forms are not consistent and the rank orders of examinees vary by forms. Equating that can adjust for difference of form difficulty may not work here. Additionally, it is important to note that raters and tasks become hidden facets in the p x f design when level scores are aggregated over raters and tasks. The variance component estimates (e.g., p x f interaction) may reflect not only form-sampling variability but also task- and/or rater-sampling variabilities.

Separate p x f G studies were also conducted for the mean scores (unrounded and rounded) averaged over raters and tasks to compare score generalizability. Table 4 reveals that variance component estimates for the person by form interaction are much smaller for the mean scores (both unrounded and rounded) than those for the current level scores. It is also important to note that raters and tasks become fixed facets when mean scores averaged over raters and tasks are used in the p x f design. Therefore, it seems that the best way to examine measurement error and generalizability is one that reflects the impacts of all possible sources of variability.

TABLE 4

Variance Component Estimates of p x f Generalizability Studies

	Forms 10cc and 11cc (Sample I)			Forms 10cc and 12cc (Sample II)		
Source	Level	Unrounded	Rounded	Level	Unrounded	Rounded
			Listening			
Person (p)	0.27151	0.26076	0.26312	0.24464	0.20623	0.25689
Form (f)	0.11221	0.09250	0.08754	0.11397	0.09820	0.09108
pf,e	0.27202	0.11369	0.17394	0.27741	0.10049	0.18420
$\hat{\sigma}(\Delta)$	0.61986	0.45408	0.51135	0.62560	0.44575	0.52467
$\hat{\Phi}$	.41	.56	.50	.39	.51	.48
			Writing			
Person (p)	0.44643	0.37236	0.40608	0.29848	0.29873	0.28367
Form (f)	0.00052	0.00208	0.00008	0.00094	0.01061	0.00953
pf,e	0.20745	0.08003	0.16590	0.23578	0.10351	0.17399
$\hat{\sigma}(\Delta)$	0.45604	0.28655	0.40741	0.48654	0.33782	0.42839
$\hat{\Phi}$	.68	.82	.71	.56	.72	.61

Note. The estimated variance components are averages over separate G studies for each rater pair.

Estimated Standard Errors of Measurement

The heavy emphasis on performance standards suggests the need to focus on standard errors and decision consistency. The Work Keys Listening and Writing scores may be used in high-stakes situations. Potential employers may use the scores in their hiring, selection, and training programs. For example, a Listening or Writing score of three on the 0-5 score scale may be determined as the level needed for an entry-level employment (i.e., a standard of proficiency). In that case, an individual with a score of three may be considered as meeting this minimum job requirement. However, due to measurement errors, the observed level score may be below or above the individual's true performance level.

Standard errors of measurement estimated in generalizability analyses can be used to compute confidence intervals containing universe (true) scores and to estimate decision consistency or uncertainty associated with measurement procedures (Cronbach et al., 1995; Linn & Burton, 1994). In G theory, there are usually two types of error variances estimated in decision (D) studies or D-study considerations: relative error variance, $\sigma^2(\delta)$, and absolute error variance, $\sigma^2(\Delta)$. Their square roots are called relative or absolute standard errors of measurement. Since the Work Keys assessment scores are used to index the level of an individual's performance (i.e., making absolute decisions), this report focuses on the estimated absolute errors-- $\hat{\sigma}(\Delta)$.

p x [(R x T):F] D-study considerations. For the measurement procedure used in the original data collection (i.e., $n_r = 3$, $n_t = 6$, and $n_f = 2$) the measurement errors are smaller for Writing (0.20 in Sample I and 0.23 in Sample II) than for Listening (0.32 in Sample I and 0.31 in Sample II). Figure 4 demonstrates that standard errors or measurement (SE) are reduced when D-study sample sizes (n'_r, n'_t, and n'_f) increase for Forms 10cc and 11cc. However, increasing the numbers of raters doesn't improve the measurement precision very much, especially for Listening, but adding more tasks and/or forms does. The patterns for Forms 10cc and 12cc Listening and Writing scores are similar to those for Forms 10cc and 11cc.

p x R x T D-study considerations. Table 5 reports absolute errors associated with different measurement procedures that contain various numbers of raters (n'_r) and tasks (n'_t). The errors are smaller for Writing scores than for Listening scores. With $n'_r = 2$ and $n'_t = 6$ the estimated measurement error is 0.26 for Writing Form 10cc and 0.25 for Form 11cc but 0.38 for both Listening Form 10cc and Form 11cc. As indicated in Figure 5, using two raters, rather than one, produce a greater decrease in measurement errors (SE) for Writing than it does for Listening. For Listening, the improvement of measurement precision is relatively small compared with the improvement reached by increasing the number of tasks. For example, consider the results for Listening Form 10cc, the $\hat{\sigma}(\Delta)$ is 0.41 with $n'_r = 1$ but 0.38 with $n'_r = 2$ when $n'_t = 6$. This means that going from one to two raters decreases $\hat{\sigma}(\Delta)$ by only about 7% with $n'_t = 6$. However, the estimate of $\hat{\sigma}(\Delta)$ is 0.46 with $n'_t = 4$ but 0.38 when $n'_t = 6$. This means that going from four to six tasks decreases $\hat{\sigma}(\Delta)$ by about 17% with $n'_r = 2$. For Writing, however, the trade-offs were similar for either going from one to two raters with $n'_t = 6$ or going from four to six tasks with $n'_r = 2$, both procedures reduce the $\hat{\sigma}(\Delta)$ noticeably (about 15%). Moreover, similar patterns were found for Form 11cc Listening and Writing.

TABLE 5

Estimated Absolute Errors and Dependability Coefficients for the p x R X T D-Study Considerations with Two Samples (I and II)

	$\hat{\sigma}(\Delta)$		$\hat{\Phi}$		$\hat{\sigma}(\Delta)$		$\hat{\Phi}$	
Task	$n'_r = 2$	$n'_r = 3$	$n'_r = 2$	$n'_r = 3$	$n'_r = 2$	$n'_r = 3$	$n'_r = 2$	$n'_r = 3$
	Listening 10cc (I)				Writing 10cc (I)			
2	0.638	0.627	0.40	0.41	0.429	0.406	0.70	0.72
4	0.455	0.446	0.57	0.58	0.311	0.293	0.82	0.83
6	0.375	0.366	0.66	0.67	0.260	0.243	0.86	0.88
8	0.328	0.319	0.72	0.73	0.230	0.215	0.89	0.90
10	0.295	0.287	0.76	0.77	0.211	0.195	0.91	0.92
	Listening 11cc (I)				Writing 11cc (I)			
2	0.654	0.642	0.40	0.41	0.400	0.381	0.69	0.71
4	0.465	0.456	0.57	0.58	0.294	0.278	0.80	0.82
6	0.382	0.373	0.66	0.67	0.250	0.233	0.85	0.87
8	0.333	0.325	0.72	0.73	0.224	0.208	0.88	0.89
10	0.299	0.292	0.76	0.77	0.207	0.191	0.89	0.91
	Listening 10cc (II)				Writing 10cc (II)			
2	0.663	0.652	0.34	0.34	0.473	0.452	0.66	0.68
4	0.471	0.463	0.50	0.51	0.345	0.327	0.78	0.80
6	0.386	0.379	0.60	0.61	0.291	0.274	0.83	0.85
8	0.335	0.329	0.66	0.67	0.260	0.242	0.86	0.88
10	0.301	0.295	0.71	0.72	0.239	0.221	0.88	0.90
	Listening 12cc (II)				Writing 12cc (II)			
2	0.602	0.585	0.36	0.37	0.424	0.404	0.57	0.59
4	0.426	0.414	0.53	0.54	0.316	0.297	0.70	0.73
6	0.349	0.339	0.63	0.64	0.270	0.251	0.76	0.79
8	0.303	0.294	0.69	0.70	0.244	0.224	0.80	0.82
10	0.271	0.263	0.74	0.75	0.226	0.207	0.82	0.85

Note. Sample I with n = 167; Sample II with n = 89.

In Sample II, the estimated absolute measurement errors for Form 12cc Listening and Writing are lower (0.35 and 0.27, respectively) than those for Form 10cc (0.39 and 0.29, respectively) with $n'_r = 2$ and $n'_t = 6$. The estimated errors for Writing are also lower than those for Listening. Again, it appears that using two raters provides more improvement over using one rater for Writing than it does for Listening. With $n'_t = 6$, going from one to two raters decreases the measurement error about 15% for Writing Form 10cc but only 5% for Listening (about 16% for Writing Form 12cc but only 7% for Listening). For Listening, using more tasks decreases measurement error more noticeably than using more raters. For example, with $n'_t = 6$, going from one to two raters, $\hat{\sigma}(\Delta)$ decreases by 5% for Listening 10cc, but with $n'_r = 2$ going from four to six tasks, $\hat{\sigma}(\Delta)$ decreases by 17%.

The estimated standard errors of measurement, $\hat{\sigma}(\Delta)$, can be used to predict the confidence intervals (or bands) in which examinees' universe (true) scores are likely to be, assuming that errors are normally distributed. For example, the 90% confidence interval containing an examinee's true performance level would be in the range of $\pm 1.645\hat{\sigma}(\Delta)$. Taking $\hat{\sigma}(\Delta)$ to be 0.382 for Listening 11cc with two raters and six tasks, the interval is about ± .628 (or 1.256).

Another use of the estimated standard error is to provide information on the rate of misclassification assuming that errors are normally distributed (Cronbach et al., 1995). For example, suppose that $\hat{\sigma}(\Delta)$ is 0.382, an examinee with a true score of 3 would have about a 10% chance of scoring below 2.5 that could be rounded to a score of 2. Similarly, about 10% of the examinees whose true score is 3 could receive a rounded observed score of 2. Test users or decision makers need to determine whether or not a 10% misclassification rate among examinees truly at 3 is acceptable.

Increasing the numbers of raters and/or tasks will narrow the sizes of the error bands or the confidence intervals and reduce the misclassification rates. As indicated in Table 6 below and Figure 6, increasing the number of tasks narrows the uncertainty ranges (or error bands) for both Listening and Writing. For example, for Writing 11cc, with $n'_r = 2$, going from two to six tasks the uncertainty range will drop from 1.315 to 0.821 score points, about a 38% reduction. In addition, increasing the number of raters from one to two will also reduce the uncertainty. For Writing Form 11cc, going from one to two raters, the number of tasks needed to get a less than one-point range of uncertainty will be dropped from six to four.

TABLE 6

Standard Errors and Error Bands with 90% Confidence for the Work Keys Listening and Writing Assessment

		Sample I				Sample II			
		10cc		11cc		10cc		12cc	
Raters	Tasks	SE	Band	SE	Band	SE	Band	SE	Band
				Listening					
1	2	0.672	2.211	0.691	2.272	0.694	2.282	0.649	2.135
1	4	0.483	1.588	0.493	1.624	0.494	1.625	0.460	1.515
1	6	0.400	1.317	0.407	1.339	0.406	1.336	0.377	1.241
1	8	0.352	1.158	0.356	1.172	0.354	1.165	0.328	1.078
1	10	0.319	1.050	0.322	1.058	0.319	1.049	0.294	0.968
2	2	0.638	2.100	0.654	2.152	0.663	2.181	0.602	1.979
2	4	0.455	1.498	0.465	1.531	0.471	1.548	0.426	1.402
2	6	0.375	1.234	0.382	1.257	0.386	1.269	0.349	1.147
2	8	0.328	1.078	0.333	1.095	0.335	1.103	0.303	0.995
2	10	0.295	0.972	0.299	0.985	0.301	0.991	0.271	0.892
				Writing					
1	2	0.490	1.611	0.451	1.484	0.530	1.744	0.480	1.580
1	4	0.360	1.184	0.340	1.118	0.395	1.299	0.366	1.206
1	6	0.304	1.001	0.293	0.965	0.338	1.111	0.320	1.052
1	8	0.273	0.897	0.267	0.879	0.305	1.005	0.294	0.966
1	10	0.251	0.827	0.250	0.823	0.284	0.935	0.277	0.910
2	2	0.429	1.410	0.400	1.315	0.473	1.555	0.424	1.396
2	4	0.311	1.023	0.294	0.969	0.345	1.136	0.316	1.038
2	6	0.260	0.856	0.250	0.821	0.291	0.957	0.270	0.887
2	8	0.230	0.750	0.224	0.737	0.260	0.854	0.244	0.801
2	10	0.211	0.693	0.207	0.681	0.239	0.785	0.226	0.745

Moreover, the impact of increasing the numbers of raters and tasks on uncertainty levels are different for Listening and Writing due to different sizes of measurement errors. For instance, increasing the number of raters has more effect on the uncertainty range for the Writing scores than for the Listening scores. For the Listening scores, increasing the number of tasks reduces the uncertainty level. However, it is still larger than one point even with $n'_t = 8$ and $n'_r = 2$. With the same numbers of raters and tasks, the Writing scores have smaller uncertainty ranges than the Listening scores.

The magnitudes of measurement errors also impact the minimum passing score and the maximum failing score needed to be confident about a pass-fail decision (Linn & Burton, 1994). If a standard of proficiency is set at three, the minimum passing score is $[3 + 1.645\,\hat{\sigma}(\Delta)]$ and the maximum failing score is $[3 - 1.645\,\hat{\sigma}(\Delta)]$ with 90% confidence in making a pass or fail decision. In other words, decision makers have about 90% confidence that examinees with scores equal to or greater than the minimum passing score will exceed the standard, and examinees with scores equal to or less than the maximum failing score will be below the standard. They have less confidence in making a pass or fail decision based on scores within the range of uncertainty. Therefore, the narrower the range of uncertainty the more precise the assessment is.

Table 7 provides the minimum passing and maximum failing scores for 90% confidence if a standard of proficiency is set at three. The values given are a function of number of tasks with $n_r = 2$. Increasing the number of tasks reduces the uncertainty of pass or fail decisions. For example, for Listening 11cc with a standard of three and $n_r = 2$, going from four to six tasks the uncertainty range between pass and fail drops from 1.531 to 1.257, about an 18% reduction. In other words, to be 90% confident, a score of 3.765 or higher is required for a pass decision with $n_t = 4$ and a score of 3.629 or higher is required with $n_t = 6$.

TABLE 7

Standard Errors and 90% Confidence Limits for Maximum Failing and Minimum Passing Scores in p x R x T D-Study Considerations

		Listening			Writing		
Form	Tasks	Error	Max. Fail	Min. Pass	Error	Max. Fail	Min. Pass
				Sample I			
10cc	2	0.638	1.950	4.050	0.429	2.295	3.705
	4	0.455	2.251	3.749	0.311	2.489	3.511
	6	0.375	2.383	3.617	0.260	2.572	3.428
	8	0.328	2.461	3.539	0.230	2.621	3.379
	10	0.295	2.514	3.486	0.211	2.653	3.347
11cc	2	0.654	1.924	4.076	0.400	2.343	3.657
	4	0.465	2.235	3.765	0.294	2.516	3.484
	6	0.382	2.371	3.629	0.250	2.589	3.411
	8	0.333	2.453	3.547	0.224	2.632	3.368
	10	0.299	2.508	3.492	0.207	2.659	3.341
				Sample II			
10cc	2	0.663	1.909	4.091	0.473	2.223	3.777
	4	0.471	2.226	3.774	0.345	2.432	3.568
	6	0.386	2.365	3.635	0.291	2.521	3.479
	8	0.335	2.448	3.552	0.260	2.573	3.427
	10	0.301	2.505	3.495	0.239	2.607	3.393
12cc	2	0.602	2.010	3.990	0.424	2.302	3.698
	4	0.426	2.299	3.701	0.316	2.481	3.519
	6	0.349	2.426	3.574	0.270	2.556	3.444
	8	0.303	2.502	3.498	0.244	2.599	3.401
	10	0.271	2.554	3.446	0.226	2.628	3.372

p x F D-study considerations. Three sets of p x f generalizability analyses were carried out using different types of scores: the current level scores and the mean scores (unrounded and rounded). The estimated standard errors were the largest for the current level scores followed by the rounded mean scores in both Listening and Writing (see Table 4). Also, the errors were smaller for Writing than for Listening when either level or mean scores (unrounded or rounded) were used. Therefore, the current level scores would have wider confidence intervals or uncertainty ranges than the mean scores for both Listening and Writing scores. For example, with a $\hat{\sigma}(\Delta) = 0.62$, the interval for the Listening level scores is ±1.02, and with a $\hat{\sigma}(\Delta) = 0.45$ the interval for the unrounded mean scores is ±0.74. Likewise, the Listening scores have wider confidence intervals than the Writing scores.

Estimated Generalizability

Generalizability analyses provide both standard errors and generalizability coefficients as indices of measurement precision. For the Work Keys Listening and Writing assessment, the interest focuses on judging examinees' levels of performance (absolute decisions), so only dependability (absolute G) coefficients are reported here.

p x [(R x T):F] design. The dependability coefficients ($\hat{\Phi}$) for this design depend, in part, upon the numbers of raters (n'_r), tasks (n'_t), and/or forms (n'_f) used in decision considerations. If only one form, two raters, and six tasks were used, the dependability coefficient would be .56 for Listening and .81 for Writing (Sample I). Figure 4 demonstrates that dependability coefficients (PHI) increase when D-study sample sizes (n'_f, n'_r, and n'_t) increase. However, increasing the number of raters beyond two doesn't improve the score generalizability substantially, especially for Listening; but adding more tasks and/or forms does. Similar patterns were observed with Sample II data. The dependability coefficients are .72 for Writing and .51 for Listening, respectively.

p x R x T design. The results of p x R x T generalizability analyses also indicated that adding more tasks increases dependability coefficients more than adding more raters, especially for Listening (see also Table 5 and Figure 5). For example, with $n'_r = 2$ and $n'_t = 6$, $\hat{\Phi}$ is .66 for Listening 11cc. With $n'_r = 3$ and $n'_t = 6$, $\hat{\Phi}$ was .67; but with $n'_r = 2$ and $n'_t = 8$, $\hat{\Phi}$ is .72. Writing scores had higher dependability coefficients than Listening scores across the three forms and the two samples of examinees. With $n'_r = 2$ and $n'_t = 6$, the range of $\hat{\Phi}$ is from .76 to .86 for Writing and is from .60 to .66 for Listening.

p x F design. As indicated in Table 4, the mean scores (both unrounded and rounded) had higher generalizability (ranging from .51 to .82 for unrounded and from .48 to .71 for rounded) than the current level scores (ranging from .39 to .68) with $n'_f = 1$ for both Listening and Writing. The Writing scores, either level scores or mean scores, are more generalizable than the Listening scores.

Conclusions

The use of the Work Keys Listening and Writing assessment needs to be accompanied by systematic evaluation of its technical qualities. The results from generalizability analyses portray some important psychometric properties about Work Keys Listening and Writing scores and may provide a better understanding of sampling variabilities of the scores and their impact on decision consistency and score generalizability.

The generalizability analyses reported here reveal that (a) examinees' scores vary from one test form to another due to large task-sampling variability, (b) the rank orderings of prompt difficulty differ across the examinees, (c) measurement errors are mainly introduced by task-sampling variability not by rater-sampling variability, (d) the Writing scores are more generalizable than the Listening scores, and (e) current level scores are less generalizable than mean scores (unrounded or rounded). However, these findings may need to be verified in future replication studies with different samples of examinees, forms, raters, and tasks. Furthermore, to fully investigate sources of sampling variation, a design with more aspects of related measurement conditions (or facets) is better than a design with hidden and/or fixed facets. In the latter, some sources of measurement error cannot be disentangled and estimated.

The finding that examinees are rank ordered differently on different forms of the Listening test suggests that conventional equating methods may not be entirely satisfactory. The result that examinees' performances vary from one task to another is consistent with other findings in performance assessments. It indicates the importance of domain specification and task sampling in test development (Shavelson, Gao, & Baxter, 1995). The assessment needs a well-defined framework specifying what to measure and a well-developed item pool to represent the designated content. The finding that one or two well-trained raters can reliably score examinees' performance is encouraging for future test operation. Moreover, the use of six prompts and two raters in the Work Keys Listening and Writing assessment leads to better generalizability than other performance assessments with fewer tasks and/or raters (Dunbar, Koretz, & Hoover, 1991).

The concern about using the current level scores generated a need for further investigation about how to convert raw scores to scale (level) scores. In addition, the small sample sizes used in the current generalizability analyses suggest that any generalization beyond the current study may not be warranted due to the possibility of large standard errors of the estimates derived from the generalizability analyses.

References

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

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of Work Keys Listening and Writing Tests. *Educational and Psychological Measurement, 55* (2), 157-176.

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.

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

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1995, summer). Generalizability analysis for educational assessments. *Evaluation Comment,* 1-29. Center for the Study of Evaluation & the National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4* (4), 289-303.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp.105-146). New York: American Council on Education & Macmillan.

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* (4), 323-342.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice, 13* (1), pp. 5-8, 15.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30* (3), 215-232.

Shavelson, R. J., Gao, X., Baxter, G. P. (1995). On the content validity of performance assessments: Centrality of domain specification. In M. Birenbaum & F. Dochy (Eds.), *Alternatives in assessment of achievements, learning process and prior knowledge* (pp. 131-141). Boston: Kluwer Academic Publishers.

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

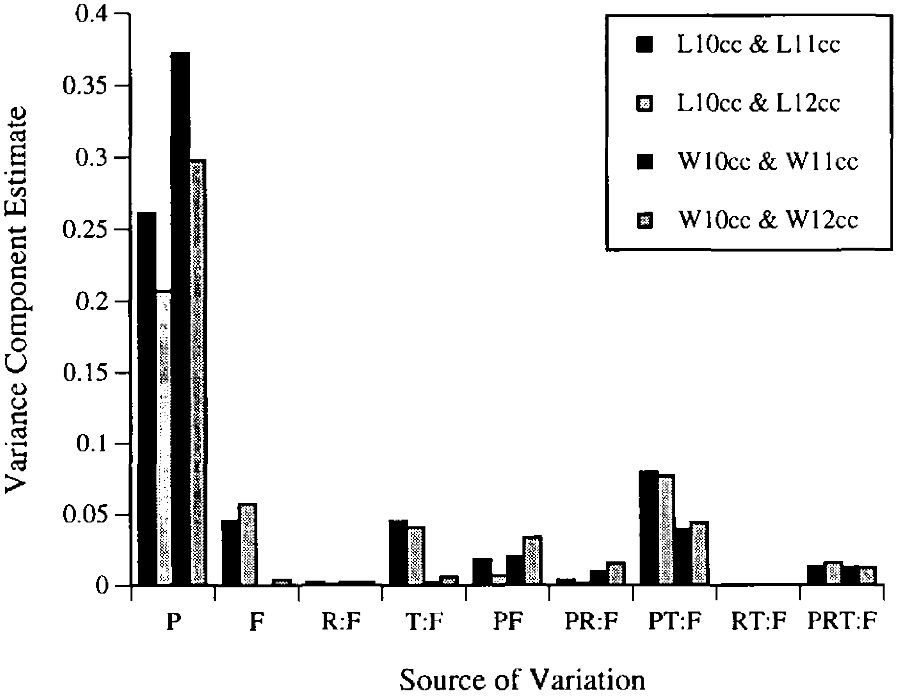

FIGURE 1. Variance component estimates for the p x [(R x T):F] design with one form, two raters, and six tasks

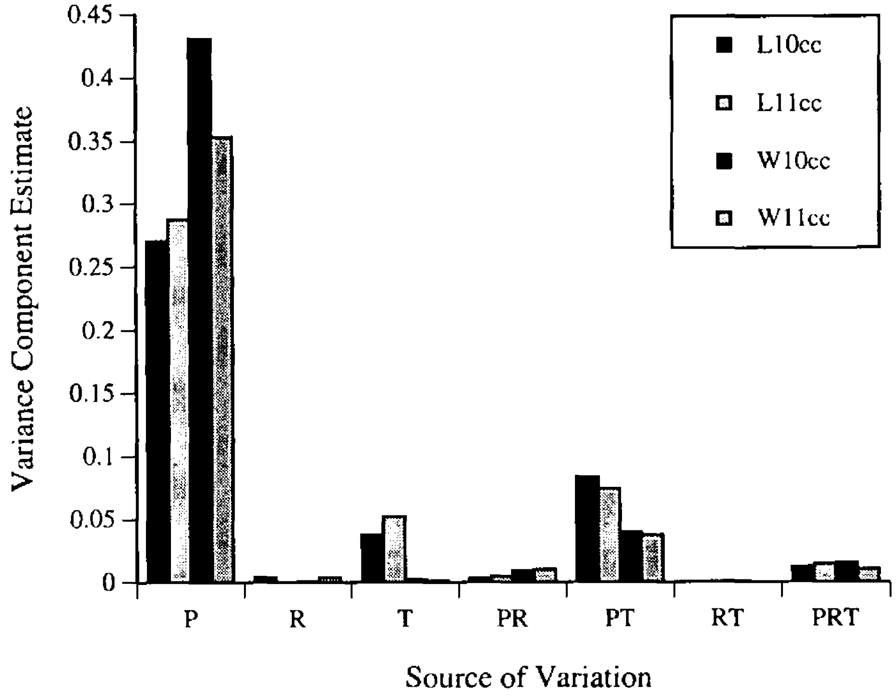

FIGURE 2A. Variance component estimates for the p x R x T design with two raters and six tasks (Forms 10cc and 11cc)

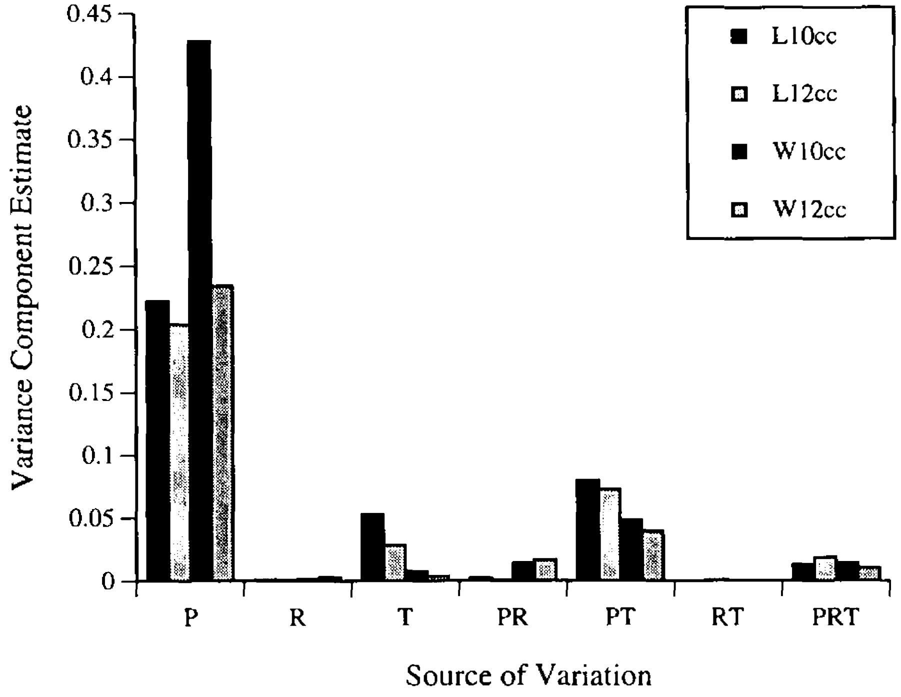

FIGURE 2B. Variance component estimates for the p x R x T design with two raters and six tasks (Forms 10cc and 12cc)

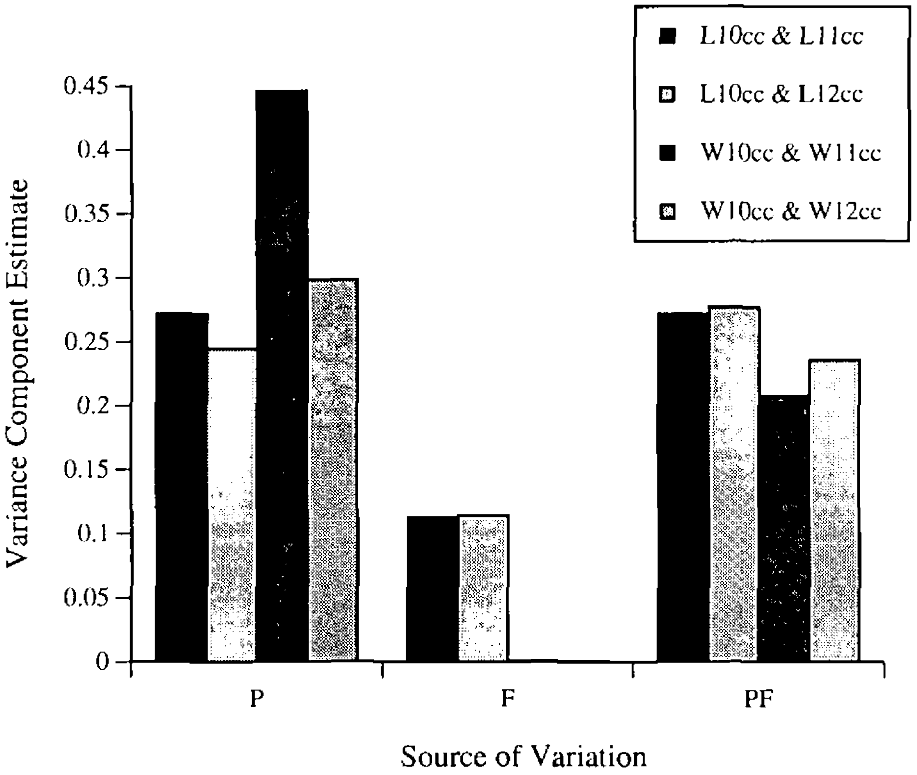

FIGURE 3A. Variance component estimates for the p x f generalizability study using corrent level scores.

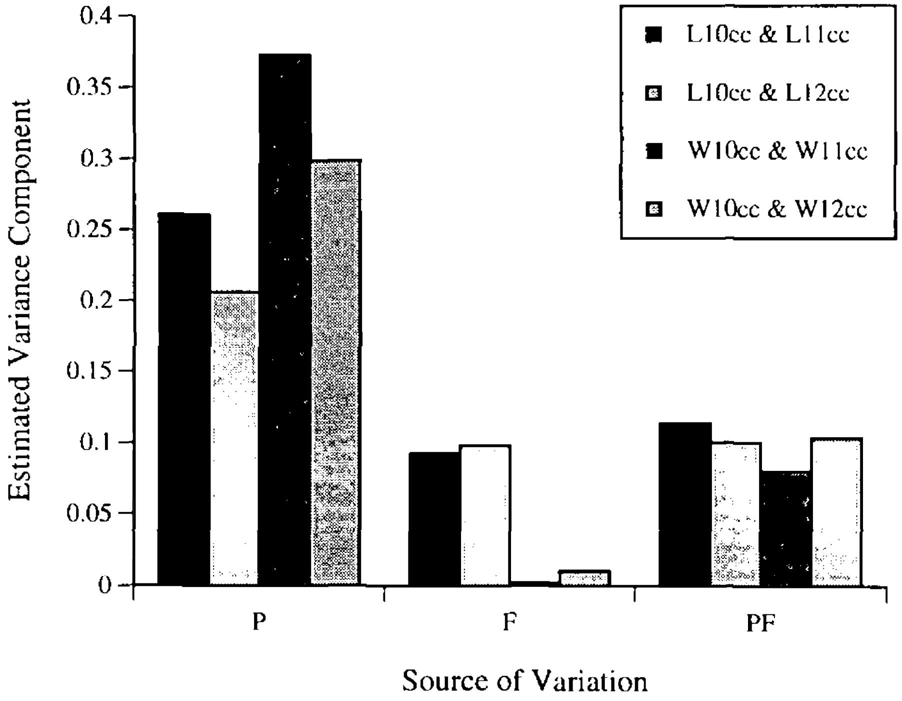

FIGURE 3B. Variance component estimates for the p x f generalizability study using unrounded mean scores.

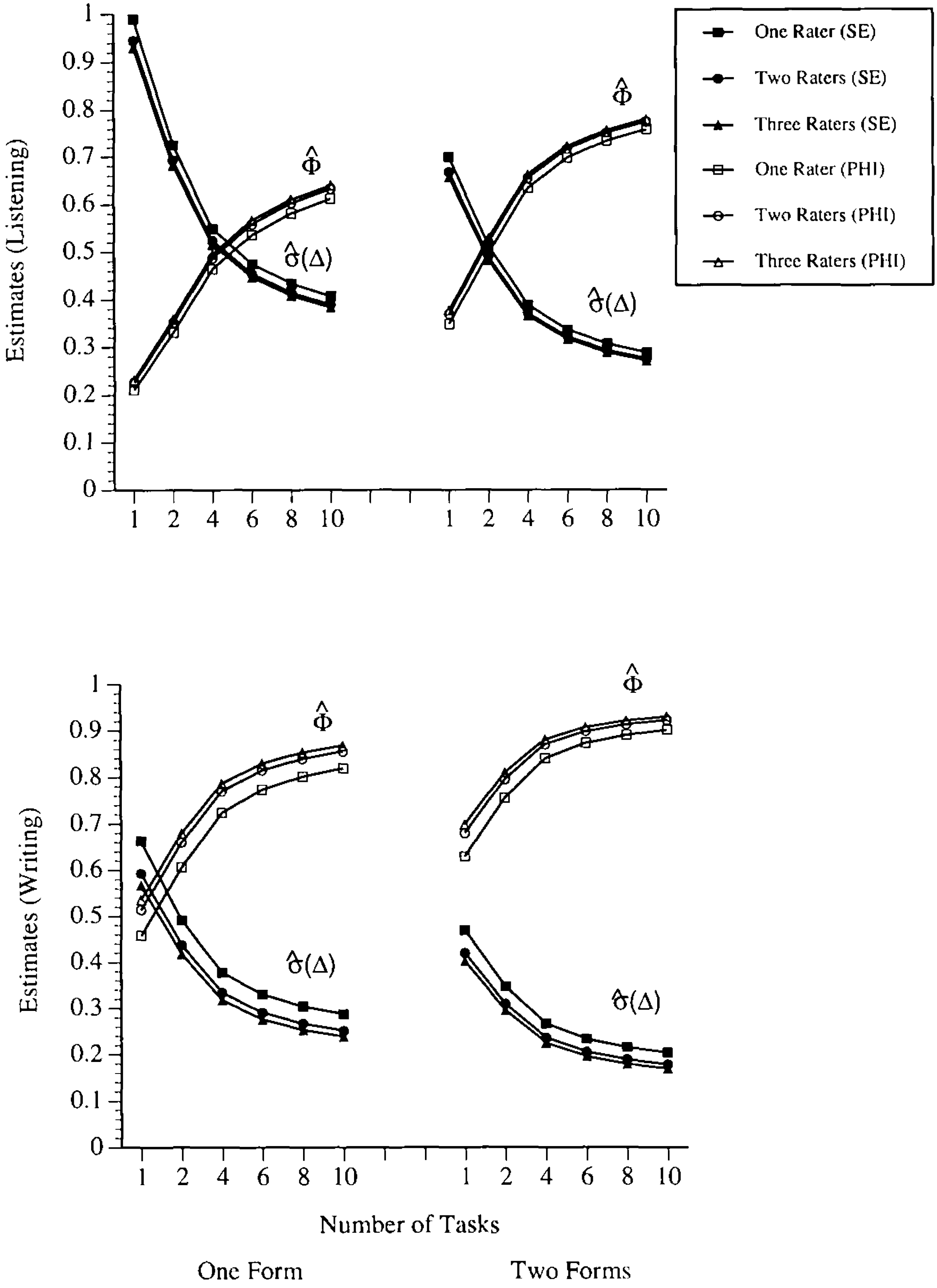

FIGURE 4. Estimated absolute error and dependability for p x [(R x T):F] D-study considerations (Forms 10cc and 11cc).

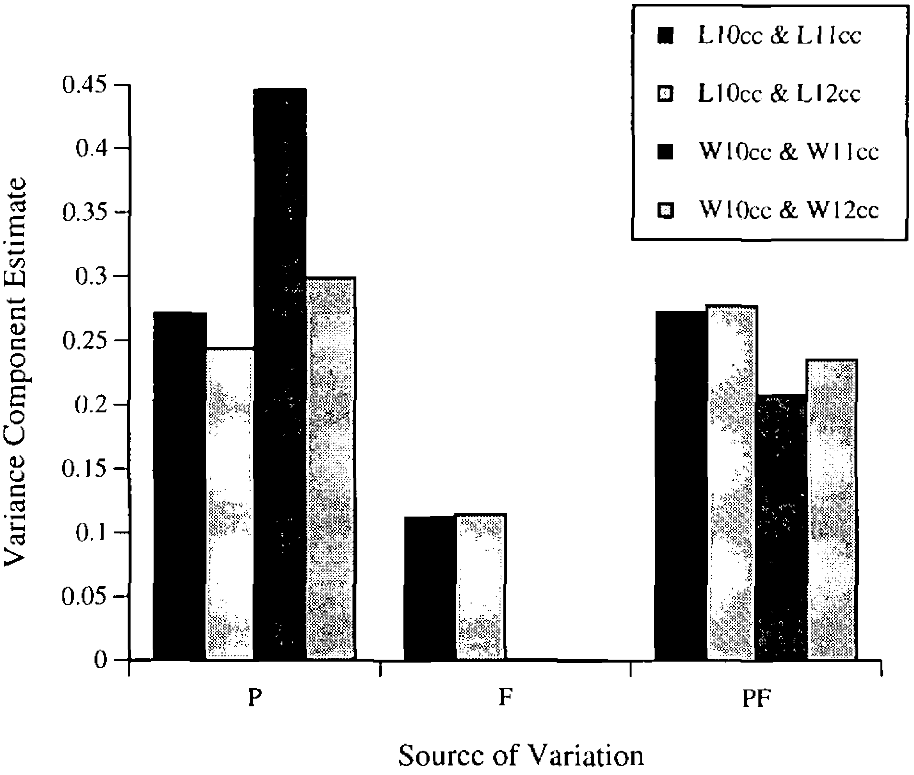

FIGURE 3A. Variance component estimates for the p x f generalizability study using corrent level scores.

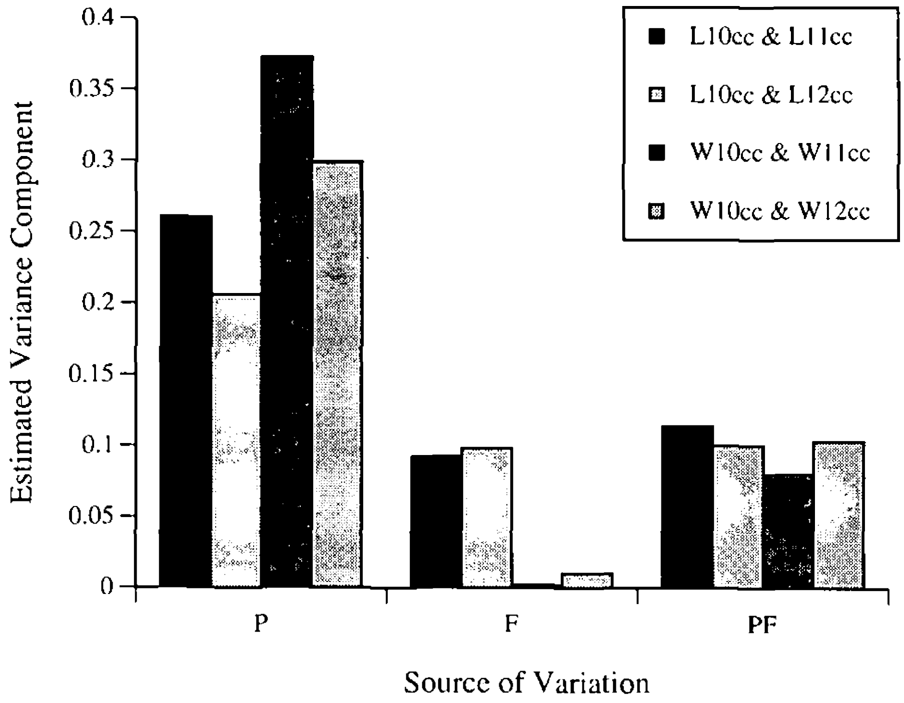

FIGURE 3B. Variance component estimates for the p x f generalizability study using unrounded mean scores.

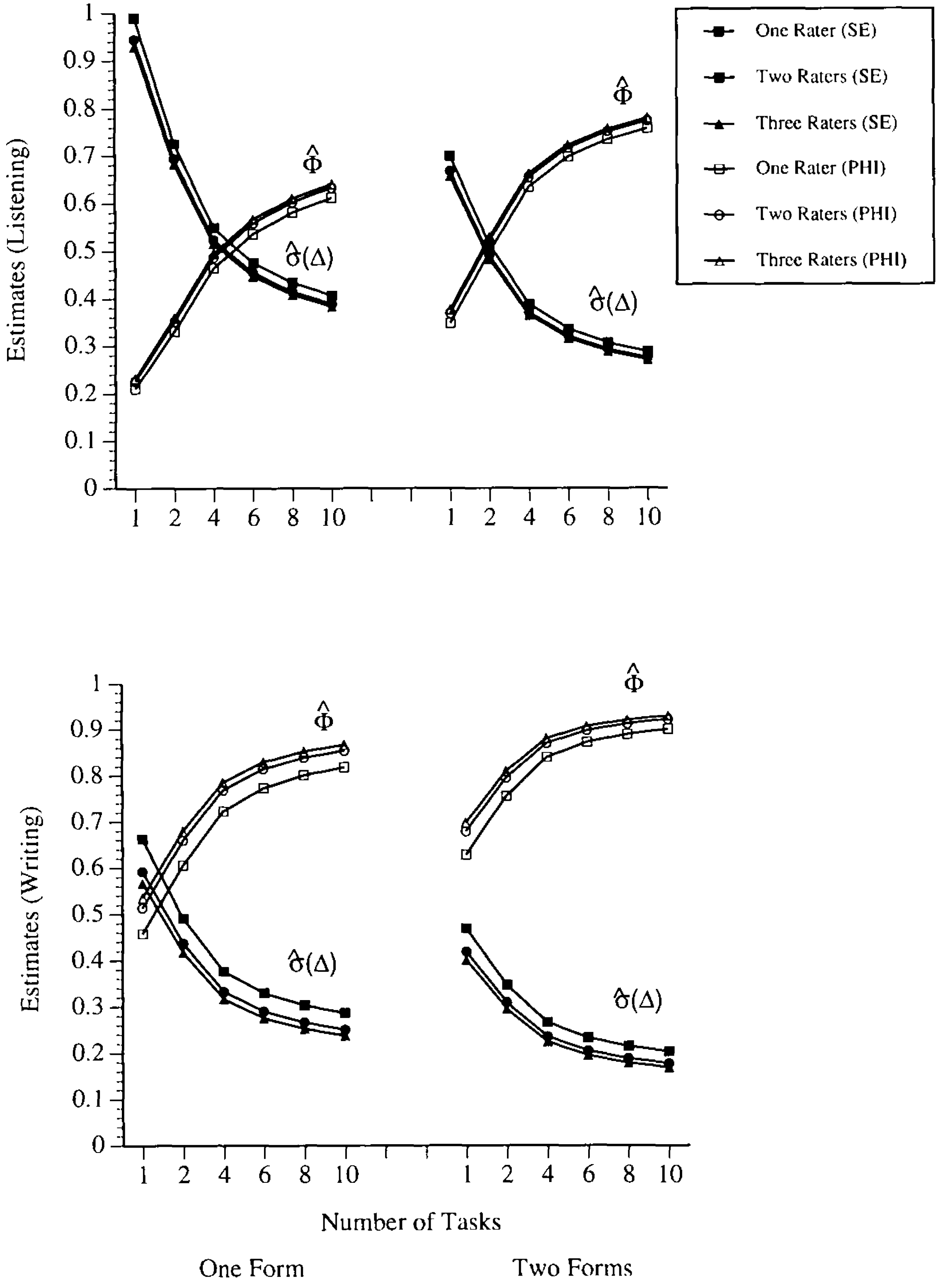

FIGURE 4. Estimated absolute error and dependability for p x [(R x T):F] D-study considerations (Forms 10cc and 11cc).

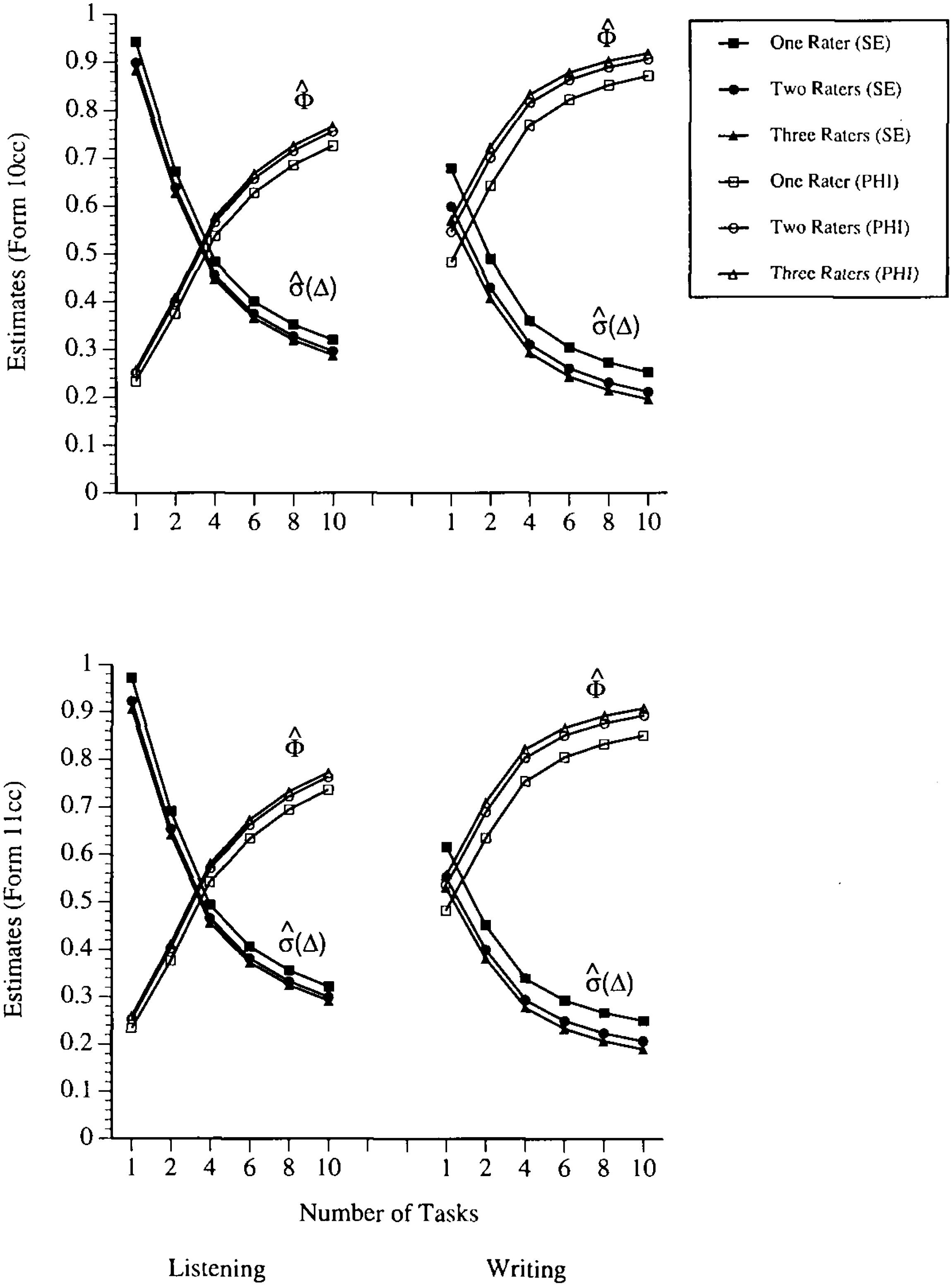

FIGURE 5. Estimated absolute error and dependability for p x R x T D-study considerations (Form 10cc and Form 11cc).

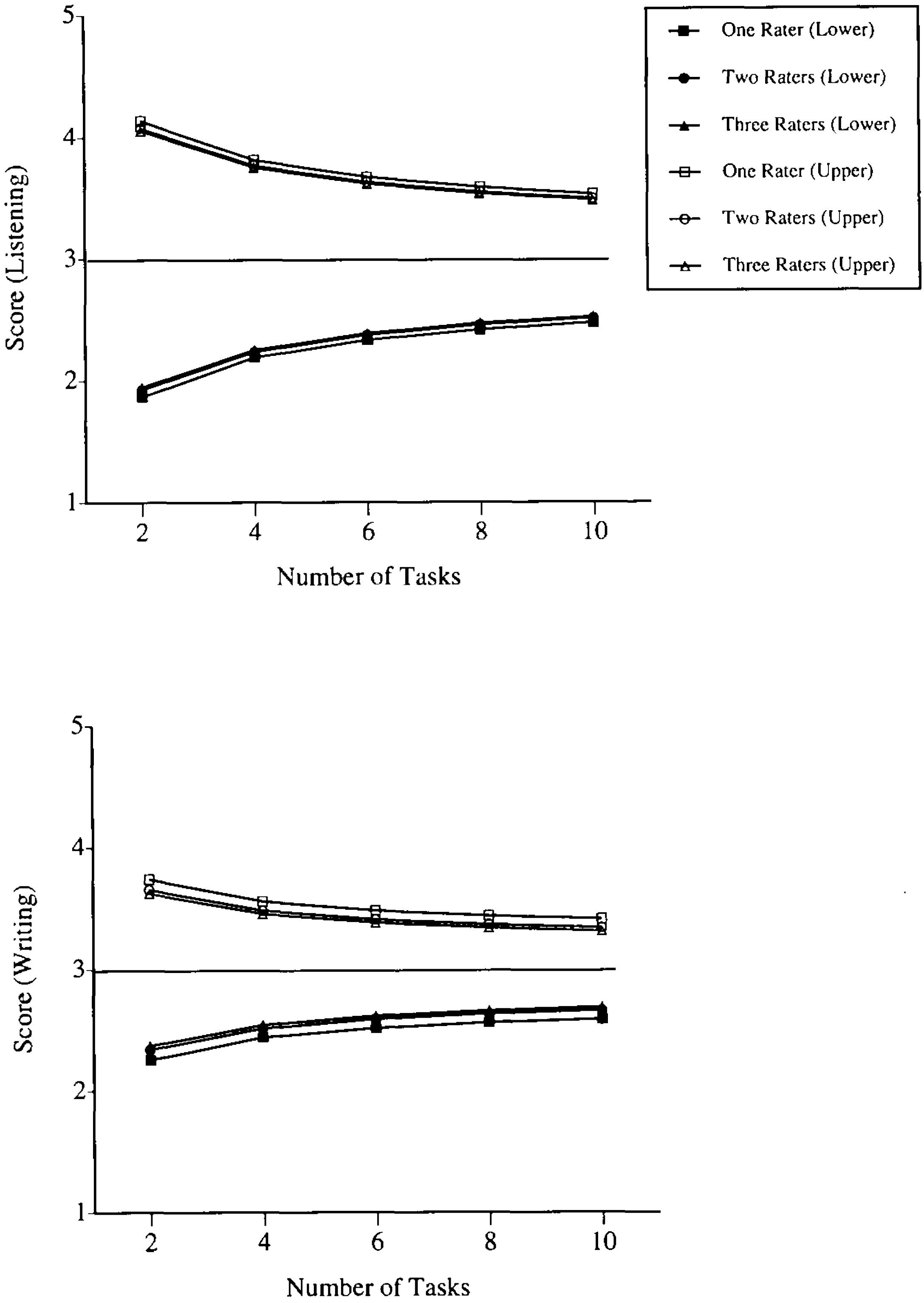

FIGURE 6. 90% confidence intervals for score of three with different numbers of raters and tasks (Form 11cc).