AL/HR-TP-1995-0036

GENERALIZABILITY THEORY AS EVIDENCE OF THE RELIABILITY AND VALIDITY OF WORK SAMPLE TESTS AND PROFICIENCY RATINGS

Kurt Kraiger

Department of Psychology
University of Colorado at Denver
P.O. Box 173364
Denver, CO 80217-3364

Mark S. Teachout

HUMAN RESOURCES DIRECTORATE
TECHNICAL TRAINING RESEARCH DIVISION
7909 Lindbergh Drive
Brooks AFB, Texas 78235-5352

December 1995

Interim Technical Paper for Period September 1993-June 1995

Approved for public release; distribution is unlimited.

ARMSTRONG LABORATORY

AIR FORCE MATERIEL COMMAND
BROOKS AIR FORCE BASE, TEXAS

NOTICE

Publication of this paper does not constitute approval or disapproval of the ideas or findings. It is published in the interest of scientific and technical information (STINFO) exchange.

When Government drawings, specifications, or the data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in anyway supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

Mark S. Teachout

MARK S. TEACHOUT, Ph.D.
Project Scientist
Technical Training Research Division

R Bruce Gould

R. BRUCE GOULD, Ph.D.
Technical Director
Technical Training Research Division

James B. Bushman

JAMES BUSHMAN, Lt Col, USAF
Chief, Technical Training Research Division

Please notify this office, AL/HRPP, 7909 Lindbergh Drive, Brooks AFB TX 78235-5352, if your address changes, or if you no longer want to receive our techinical reports. You may write or call the STINFO office at DSN 240-3853 or commercial (210) 536-3853.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
Dccembcr 1995

3. REPORT TYPE AND DATES COVERED
Interim Paper - September 1993 - June 1995

4. TITLE AND SUBTITLE
Generalizability Theory as Evidence of the Reliability and Validity of Work Sample Tests and Proficiency Ratings

5. FUNDING NUMBERS
PE - 62205F
PR - 1121
TA - 12
WU - 00

6. AUTHOR(S)
Kurt Kraiger
Mark S. Teachout

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Psychology Department
University of Colorado at Denver
P.O. Box 173364
Denver, CO 80217-3364

8. PERFORMING ORGANIZATION REPORT NUMBER

9. Box SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Armstrong Laboratory
Human Resources Directorate
Technical Training Research Division
7909 Lindbergh Drive
Brooks AFB, TX 78235-5352

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AL/HR-TP-1995-0036

11. SUPPLEMENTARY NOTES
Armstrong Laboratory Technical Monitor: Dr. Mark S. Teachout, DSN: 240-2932
Comm: (210) 536-2932

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release: distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This paper uses generalizability (G) theory as a framework to investigate the reliability and construct validity of the Air Force Job Performance Measurement System. G theory was a useful technique for examining the psychometric quality of the measurement system because it permits the specification and estimation of multiple sources of measurement error. G theory was applied to newly developed work sample procedures, Walk-Through Performance Tests (Hedge & Teachout, 1992) and to a series of job proficiency ratings comprised of four different rating forms collected from three rating sources. The results provided evidence of strong convergent and discriminant validity of these work sample tests, strong convergent validity across rating forms, and moderate discriminant validity of the rating system. However, ratings did not generalize across self, peer and supervisor rating sources. Results are discussed in terms of their practical and theoretical implications.

14. SUBJECT TERMS
Analysis of Variance
Convergent Validity
Construct Validity
Discriminant Validity
Generalizability Theory
Hands-on Testing
Interview Testing
Job Performance Measurement
Multi-Trait Multi-Method (MTMM)
Performance Ratings
Reliability
Walk-Through Performance Tests
Work Sample Tests

15. NUMBER OF PAGES
25

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION O ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED

CONTENTS

LIST OF TABLES

PREFACE

This paper demonstrates the usefulness of generalizability theory as a data analysis procedure and summarizes key results from studies conducted as part of the Air Force Job Performance Measurement project. The levels of reliability established for the rating and work sample criteria described in this study make them useful for training evaluation purposes.

Portions of this research were completed under prime contract number F41689-84-D-0001 with Universal Energy Systems for the Training Systems Division of the Air Force Human Resources Laboratory, Brooks Air Force Base, TX. This paper was completed under in-house Work Unit No. 1121-12-00. A previous version of this paper was presented at the annual meeting of the Society for Industrial and Organizational Psychology, St. Louis, MO, April, 1991. The authors are grateful to Herman Aguinis for his helpful comments on an earlier version of this manuscript.

GENERALIZABILITY THEORY AS EVIDENCE OF THE RELIABILITY AND VALIDITY OF WORK SAMPLE TESTS AND PROFICIENCY RATINGS

SUMMARY

Investigations of construct validity require an accumulation of evidence consistent with a priori expectations about variables related to, and unrelated to, the construct of interest. In the current investigation, generalizability (G) theory was used as a framework for an investigation of the construct validity of a Job Performance Measurement System. Since G theory permits the specification and estimation of multiple sources of error, it was a useful mechanism for examining the psychometric quality of the measurement system. G theory was applied to both Air Force Walk Through Performance Tests and job proficiency ratings. The results provided evidence of strong convergent and discriminant validity of the work sample tests, strong convergent validity over rating forms, and moderate discriminant validity of the rating system. However, ratings did not generalize over rating sources. Practical and theoretical implications of the results are discussed.

I. INTRODUCTION

Generalizability (G) theory was developed by Cronbach and his associates as an alternative to classical test theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). To date, there have been relatively few applications of G theory to the domain of performance measurement. In one, Kraiger and Teachout (1990) used G theory analyses to provide construct-related evidence of the validity of proficiency ratings of Air Force jet engine mechanics. In this study, we use G theory to investigate the reliability and construct validity of both proficiency ratings and Walk-Through Performance Testing (WTPT) scores. Further, we extend the analyses to a total of eight Air Force occupational specialties, permitting the comparison of results across jobs. We first provide a brief overview of G theory, then discuss G theory as evidence for construct validity, and finally present results for the performance measures collected across the eight specialties.

Generalizability Theory

G theory was developed as a multi-faceted framework for examining the dependability of behavioral measures. While classical test theory implicitly recognizes multiple error sources (e.g., raters or testing occasions), it requires independent research designs to generate appropriate indices for each source (e.g., inter-rater or test-retest reliability coefficients). Thus, relationships among different types of measurement error are unclear and inestimable. In contrast, G theory explicitly permits simultaneous estimation of multiple sources of error within a single design.

G theory assumes that a measurement taken on a person represents a sample of behavior; we are interested in that sample only because it informs us about the person's behavior within a much broader context, or universe of generalization. Thus, a performance rating on a single dimension of performance is of interest only to the extent that it informs us about the performance of the ratee on other dimensions, on other occasions, or by other raters. G theory replaces the classical notion of reliability with one of generalizability, which addresses how well an observed score generalizes to inferences about behavior in the universe of generalization.

G theory investigations include both generalizability (G) studies and decision (D) studies. G studies are typically conducted on raw scores in order to estimate the contribution of measurement conditions to total observed score variance. In a typical G study, the researcher specifies a set of facets (or factors) which may affect the variability of scores. For example, in a performance rating study rating sources, rater training, or rating purpose might be facets for study. The researcher than obtains scores on a measure while sampling from multiple conditions within each specified facet. For example, if rating purpose was a facet, scores might be sampled when the ratings were collected for administrative, feedback, or research purposes. G study analyses then partition total score variance into that due to each facet, each interaction among facets, and individual differences. Calculated variance components in a G study represent estimated variance about universe scores (i.e., true scores) for single observations (e.g., an average person evaluated on a single item by a single rater). A G coefficient can be calculated as the ratio of universe score variance to observed score variance; this G coefficient represents the proportion of observed score variance attributable to individual differences.

D studies often use G study variance components as data, and are performed to estimate the dependability of a measure under a specific set of measurement conditions, or to predict the measurement conditions necessary to achieve a desired level of reliability. . A G coefficient calculated during a D study indicates the reliabililty of the measure under a specific set of measurement conditions. D studies can also be used to study other aspects of decision-making using the measure. For example, the reliability of the measure when making absolute vs. relative decisions can be investigated. The relationship between G and D studies is illustrated by the following example. In a job analysis project, a G study may be conducted to estimate the total variance in task importance ratings attributable to raters' job level, or physical location. Using the G study variance component for locations, a D study can be conducted to estimate the number of locations which must be sampled to achieve reliable ratings. Brennan (1983) and Shavelson and Webb (1991; Shavelson, Webb, & Rowley, 1989) provide more extensive treatments of G theory.

G Theory as Construct-Related Evidence of Validity

Evidence of reliability (using G theory) serves as preliminary support for establishing the construct validity of performance measures. In addition, G theory can support inferences of construct validity through the formulation and testing of expectations about the relative size of variance components (J. P. Campbell, 1976; M. T. Kane, 1982; Kraiger & Teachout, 1990). Estimated variance components reflect the magnitude of error in generalizing from a person's score under certain measurement conditions to his or her true score.

Construct validation. The process of construct validation was originally defined as a multi-step process of defining and testing a nomological network, where the net contained a series of hypotheses about the measure of interest and other measures or experimental factors (Cronbach & Meehl, 1955). A similar approach is still advocated (e.g., Austin, Villanova, J. S. Kane, & Bernardin, 1991). However, for many applied researchers, the fundamental requirements for construct validation were defined by D. T. Campbell and Fiske (1959): Any investigation of the construct validity of a measure should reveal that measure has convergent validity, discriminant validity, and relatively low method bias. D. T. Campbell and Fiske also argued that these three requirements could be tested fully only in a multitrait-multimethod design. D. T. Campbell and Fiske defined each requirement, both conceptually and operationally in terms of the relative size of correlations in a multitrait-multimethod matrix. For example, convergent validity was defined as the extent to which independent methods agreed in assessing a particular trait; operationally, convergent validity was revealed by validity diagonals which were significantly different than zero.

Applying the multitrait-multimethod analysis to performance ratings, Kavanagh, MacKinney, and Wolins (1971) noted two important shortcomings of D. T. Campbell and Fiske's (1959) seminal work. First, judging the relative size of correlations across blocks of a matrix required a high degree of subjectivity. Second, the number of calculations and comparisons become overwhelming with even moderately large matrices.

Analysis of Variance Approach. Drawing on work by Stanley (1961), Boruch, Larkin, Wolins, and MacKinney (1972), and others, Kavanagh et al. (1971) applied Analysis of Variance (ANOVA) to the analysis of multitrait-multimethod matrices. As noted by these researchers, the principal difference between a two-factor ANOVA design and a measurement design is that in the latter case, one of the factors is the sample of persons being measured. Thus, just as ANOVA can be used in a true experiment to partition variance due to treatments and variance due to treatments and contingency factors, the same technique can be used in a measurement design to partition variance due to persons and variance due to the interaction of persons with traits, sources, etc. Kavanagh et al. provided formulas for partitioning total observed score variance into multiple sources, e.g., variance due to individual differences (i.e., ratees) or variance due to the interaction of ratees and methods (or rating sources).

G Theory Approach. As suggested in this manuscript, generalizability theory can also be used to analyze multitrait-multimethod matrices, as well as more sophisticated designs. Thus, it is important to consider the relationship between G theory and the ANOVA approach. G theory in fact uses ANOVA to estimate variance components, so that for some designs, both approaches would yield equivalent results. However, the ANOVA approach is best thought of as a special case of G theory. G theory is a more generalized approach than ANOVA in several ways. G theory is more easily adaptable to more complex designs (e.g., multiple facets, or nested designs. ANOVA analyses are restricted to observed (person) score variance, while G theory can be applied to observed score variance or total score variance if absolute decisions are to be made (e.g., during criterion-based testing). Most importantly, in the ANOVA approach, method factors are considered fixed (Kavanagh et al., 1971), while in G theory, these factors can be treated as random or fixed. An important implication of treating factors as random is that more variance components can be estimated. For example, in Kavanagh et al.'s analysis of ratings of managers on multiple traits by multiple sources, only four variance components were estimated. Using G theory, seven variance components could have been calculated.

Thus, while G theory and ANOVA can both be used to analyze certain designs, G theory can be used in a greater number of contexts, and also provide more data in each analysis. It is also important to recognize that the interpretations we make of estimated variance components differs from those of Kavanagh et al. (1971). These differences, described below, arise not from variations in data analysis methods, but from different operationalizations of the desideratum for validity outlined by D. T. Campbell and Fiske (1959).

Two major requirements for construct validity are convergent and discriminant validity (D. T. Campbell & Fiske, 1959; Cronbach & Meehl, 1955). Convergent validity is evident when there is a substantial relationship between scores on the focal measure and another measure of the same construct. D. T. Campbell and Fiske suggested that evidence for convergent validity should come from "entries in the validity diagonal ... significantly different from zero (pg. 82)." In G theory, such evidence occurs when scores are invariant over conditions of a facet (M. T. Kane, 1982), provided conditions in the facets are different operationalizations of the construct (e.g., multiple measures, multiple raters, etc). Thus, convergent validity is evident when persons are similarly ranked over methods. Boruch et al. (1962) proposed a similar interpretation, but Kavanagh et al. suggested that convergent validity should be interpreted from large values for s_p^2 (universe score variance, or variance in persons across methods, sources, etc.). We see two problems with this interpretation. First, for designs with multiple methods facets (e.g., sources and forms), it is impossible, using s_p^2, to isolate the convergent validity of either method. Secondly, at an operational level, the index seems inconsistent with Campbell and Fiske's definition of convergent validity since it is based on the average correlation within heterotrait-heteromethod blocks, not the validity diagonal (see Kavanagh et al., Table 5).

Discriminant validity is evident when persons are differentially ranked on measures which assess different attributes. For example, one would not necessarily expect high correlations between measures of verbal knowledge and creativity. When the measures to be compared are traits on a single instrument (e.g., two dimensions on a rating form), it is important to consider whether a single or multiple construct(s) underlie the measurement domain. In the case of achievement testing (e.g., final course exams), it is often assumed that a single construct accounts for test performance, and persons should be similarly ranked across test items. However, in the case of performance measurement, the construct domain is usually conceptualized as multi-dimensional in nature (Bernardin & Beatty, 1984). Thus, it is worthwhile to investigate the degree to which persons are differentially ordered by tasks on the WTPT, or by rating dimensions on the proficiency rating forms. In G theory, discriminant validity can be estimated by the variance component s_{px}^2, where x is either tasks or dimensions. Larger values for the variance component denote discriminant validity. A similar interpretation holds for both G theory and ANOVA approaches.

The G theory and ANOVA approaches do differ in their interpretations concerning evidence for method bias. Campbell and Fiske (1959) defined method bias, or halo, as "systematic variance among...scores...due to responses to measurement ... factors (pg. 81)." Kavanagh et al (1971) interpret large values $\underline{s}_{pm}^2$ as evidence of method bias, while we interpret small values of the same index as evidence of convergent validity. This incongruity is the result of fixing traits in the ANOVA approach. Given that traits are fixed, inferences of validity can only be made at the total score level. In G theory, with traits (or items) treated as random, method bias can be directly assessed at the item level through inspection of $\underline{s}_{ps(t:m)}^2$. This value would be large if the covariance among traits was greater for some sources than for others. Again, we believe this interpretation is more closely aligned with the definition offered by Campbell and Fiske (1959).

In the studies described in this paper, we applied generalizability theory to WTPTs and job proficiency ratings collected as part of the Air Force Job Performance Measurement (JPM) system. Consistent with the approach we outlined above, we used G theory to provide evidence of the reliability and construct validity of these job performance measures.

II. METHODS

Participants

Personnel from eight Air Force enlisted specialties participated in this research as part of a large-scale effort to develop criterion measures for the validation of selection and classification tests and the evaluation of training programs (Hedge & Teachout, 1986). The tested specialties and associated sample sizes are: Jet Engine Mechanic (n=255), Air Traffic Control Operator (n=172), Avionic Communications Specialist (n=98), Information Systems Radio Operator (n=158), Aircrew Life Support (n=216), Personnel Specialist (n=218), Precision Measurement Equipment Laboratory Specialist (n=138), and Aerospace Ground Equipment Mechanic (n=264).

Measures

Data were collected on hands-on and interview work sample tests and four types of rating forms. The job content for all measures was identified through an extensive task sampling plan which included information on the tasks performed, the relative amount of time spent performing these tasks, and task learning difficulty (Lipscomb & Dickinson, 1988).

Walk-Through Performance Testing. Walk-Through Performance Testing combines an interview format with a more traditional hands-on work sample test approach to provide a task-level measure of individual technical job competence. Details about the development and content of the WTPT measures are provided in Hedge and Teachout (1992). In brief, the measures require the examinee to describe or perform tasks at the work setting under the observation of a trained test administrator. The test administrator records on a checklist whether the step was described or performed correctly. Test administrators were active-duty or recently retired or separated experts in the specialties tested. They received one to two weeks of training in observation and evaluation, interviewing, scoring, and WTPT procedures. (See Hedge & Teachout, 1992 for a summary of the administrator training, as well as evidence of the reliability and accuracy of administrator scoring).

Rating Measures. Four rating forms were developed using the same job analysis information used for the WTPT. Ratings were collected from supervisors, peers, and job incumbents on a 5-point anchored rating scale, ranging from 1 (never meets acceptable level of proficiency) to 5 (always meets acceptable level of proficiency).

The task-level rating forms used graphic rating scales to measure technical proficiency on tasks representative of the job content domain. The number of task scales varied from 25 to 40 across jobs. The dimensional-level rating forms were developed to measure technical proficiency on four to 10 dimensions (depending on the specialty). Behavioral descriptors were developed by subject matter experts for each of the five scale values, using a behavioral summary statement approach (Borman, 1979). An Air Force-wide rating form assessed eight general performance factors required for success in all Air Force jobs (e.g., technical knowledge/skill, knowledge of and adherence to regulations/orders). Finally, a global rating form consisted of two scales intended to measure ratees' overall technical and overall interpersonal proficiency.

Procedure. In a group orientation session, the research project was described, and raters were familiarized with the measures used in the project. This orientation was followed by one hour of frame-of-reference and rater error training with content adapted from McIntyre, Smith, and Hassett (1984). The work sample testing occurred over several days at each site. Each incumbent was tested individually by a test administrator.

Analyses. The data were analyzed using GENOVA, a Fortran-based computer program designed for generalizability analyses (Crick & Brennan, 1982). D study analyses were conducted with G study estimated variance components as input. The number of conditions observed for each facet were systematically varied at the D study level to estimate generalizability under measurement conditions of various levels of practical interest and complexity. For example, generalizability coefficients were computed for the multiple combinations of WTPT scores (e.g., one method using five ten-step tasks, or two methods each with ten 15-step tasks). Operationally, a D study variance component is computed by dividing the G study variance component by the number of conditions of any facet indicated by its subscript. For example, $\underline{s}_{mt}^2$ would be divided by 24 if incumbents completed 12 tasks on each of two methods.

III. DESIGNS AND RESULTS

Walk Through Performance Tests

Generalizability Designs. Three facets were considered for investigating the dependability of WTPT scores. The first facet was the assessment method, with hands-on and interview components as the conditions of the facet. The second facet was comprised of the tasks (i.e., constructs) that were measured by both the hands-on and interview components. Typical WTPTs consisted of 20-25 tasks. For each specialty, these tasks can be considered random samples of a larger possible universe of tasks which could comprise the WTPT.

There were three types of tasks included in the WTPT: Overlap tasks common to both the hands-on and interview components, tasks unique to the hands-on component, and tasks unique to the interview component. Thus, overlap tasks were assessed by both methods, while unique tasks were assessed by one WTPT method but not the other. For purposes of analysis, there were two possible generalizability designs for investigating variance due to tasks. One analysis included only the overlap tasks and treated tasks as crossed with methods, since each task is assessed by each method and each method includes all tasks. A second analysis included only unique tasks and treated tasks as nested within methods since tasks differed for each method of the WTPT. To maximize the number of tasks analyzed (and reduce sampling error), analyses were conducted with both common and unique tasks nested within methods. For example, eight unique tasks and six common tasks may have been analyzed as nested within a method even though these common tasks were not actually nested. For this paper, only the results for the latter nested design are presented, since analyses of both designs yielded similar results.

The final facet of interest was the number of items or steps comprising individual tasks on the WTPT. Items are nested within tasks since they were different for each task. When WTPT scoring procedures were established, steps were identified which were either the most observable, or those that captured the essence of the task. Further, while Air Force procedural manuals describe a preferred sequence of task steps, these are often only one of several ways in which incumbents can (and do) accomplish the tasks. Thus, the items can be considered random samples of larger possible universes of possible items for each task. Different individuals may leave out or add steps, or perform steps in different sequences. The items facet for the WTPT was unbalanced since the number of steps for a task ranged from as little as four to over 30. To balance the items facet (and avoid biased mean square estimates, see Searle, 1971), tasks with only a few items were dropped from the analyses, and items were randomly selected from longer tasks to match the number of items in the shortest remaining tasks.

Methods, tasks, and items were each treated as random facets for purposes of analysis. In G theory, any facet may be considered random if there is at least one other condition not represented in the design which could be meaningfully substituted for existing conditions (Brennan, 1983; Shavelson & Webb, 1991). For example, performance ratings or job knowledge tests are alternative methods for assessing Airmen proficiency.

Results Supportive of Construct Validity. Table 1 shows the expected results for variance components interpretable as evidence of the construct-related validity of the measures. The first two expectations are for WTPT variance components. Note that the use of "small" and "large" are somewhat arbitrary since the absolute values of G study variance components depend on the total variance and number of factors in a design. We use the labels of small and large to refer to the size of variance components relative to other variance components within the design and across specialties.

Table 1. Expected Results Supportive of Construct Validity.

	s^2	Expected Magnitude	Type of Evidence
1	s_{pm}^2	Small	Convergent Validity, WTPT: Scores are invariant over WTPT methods (the interview was developed as a surrogate for the hands-on test)
2	$s_{p(t:m)}^2$	Large	Discriminant Validity, WTPT: Persons are differentially ranked by WTPT tasks; individuals vary in task experience and capacities to perform various tasks
3	s_{pf}^2	Small	Convergent Validity, Rating Forms: Scores are invariant over rating forms since, at the total score level, each assesses total job performance.
4	$s_{p(i:f)}^2$	Large	Discriminant Validity, Rating Forms: Persons are differentially ranked by rating scales; individuals vary in task experience and their capacities to perform various on-the-job tasks.
5	s_{ps}^2	Large	Low Convergent Validity, Rating Sources: Persons are differentially ranked by sources; different sources may observe different performance samples, or impute different percepts when forming judgments of effectiveness

G Study Results. Results of the G study analyses across specialties are presented in Table 2. The estimated variance components indicated the contribution of each effect to total score variance. The $s_{p(i:t:m)}^2$ variance component is a residual term and represents variance due to both random error and the interactions of items nested with tasks which are nested within methods. Since the interaction effect and random error are confounded in this term, neither can be interpreted. Since other error variance terms are small, it is likely that the larger values for $s_{p(i:t:m)}^2$ indicate random error within scores.

However, it is instructive to look at the relative sizes of variance components both across and within specialties. Consistency in the magnitude of variance components permits inferences about the extent to which similarly-sized variance components would be found in other specialties or studies. For example, values in Table 2 for $\underline{s}_{m}^{2}$ are very similar across specialties, indicating that both methods yielded similar mean scores, regardless of the specialty examined. This consistency in results over specialties suggests that similar results would be found if the WTPT were applied to additional specialties. In contrast, values for $\underline{s}_{t:m}^{2}$ are small in some specialties, but larger in others. Thus, the extent to which overall task scores will vary by method of assessment depends on the specialty tested, and the results of these analyses are not easily generalized to other untested specialties.

Inspection of variance components within specialties reveals the relative contributions of person effects to total variance, and it is these results which are used to assess construct validity. Consistent with expectations shown in Table 1, the obtained values for $\underline{s}_{pm}^{2}$ were very low[1] and near zero in most specialties (indicating scores for incumbents converged across methods). This indicated that persons were similarly ordered regardless of evaluation methodology. Thus, the interview is a suitable surrogate for the hands-on format (cf., Hedge & Teachout, 1992).

Also consistent with expectations, the $\underline{s}_{p(t:m)}^{2}$ component was relatively large. The results in Table 2 reveal that this value was one of the three largest variance component in all specialties. Excluding the residual term (which is uninterpretable), the $\underline{s}_{p(t:m)}^{2}$ variance component was largest or second largest term in all specialties. The stability of this index can be assessed by determining the ratio of the variance component to observed score variance and comparing this value across specialties. This value ranged between .101 for Precision Measurement Equipment Laboratory Specialists and .325 for Information System Ratio Operators; most values fell between .100 and .200. This suggests that approximated 10 to 20% of the variance in WTPT scores can be attributed to discriminant validity across job tasks, regardless of specialty. Persons are differentially ordered by tasks.

D Study Results. G coefficients were computed for a variety of possible values for each facet in the design. A generalizability coefficient represents the proportion of observed score variance attributable to universe score variance or individual differences, and indicates the dependability of a measure under a particular set of conditions. Because of the relative size of several G study variance components, G coefficients were significantly affected by averaging scores over multiple tasks and both methods. Increasing the number of steps on each task had only a negligible effect on the generalizability of scores

[1] Due to sampling error, calculated values for estimated variance components may sometimes be less than .00. Since variance components must be greater than or equal to .00, estimated values less than .00 should be interpreted as zero. It is only when a large proportion of calculated values are less than .00 that either the design or the data should be questioned (Brennan, 1983).

Table 2. Estimated Variance Components for G Study of Walk-Through Variables With Tasks Nested Within Methods.

				Job:				
	JEM	ACS	ATC	ISRO	ALS	PS	PMEL	AGE
Effect	$\underline{s}^2$	$\underline{s}^2$	$\underline{s}^2$	$\underline{s}^2$	$\underline{s}^2$	$\underline{s}^2$	$\underline{s}^2$	$\underline{s}^2$
p	.008	.013	.007	.029	.018	.038	.004	.011
m	.013	.001	-.001	-.001	.004	-.007	.004	-.006
t:m	.003	.014	.012	.008	.026	.013	.010	.036
i:t:m	.020	.030	.032	.009	.037	.008	.037	.053
pm	.001	-.001	-.002	-.003	-.001	-.031	-.001	-.006
p(t:m)	.019	.032	.018	.051	.027	.051	.011	.037
p(i:t:m)	.144	.108	.128	.080	.119	.078	.095	.126
G Coeff.	.567	.901	.521	.936	.933	.951	.853	.869

Note. JEM = Jet Engine Mechanic, ACS = Avionic Communications Specialist, ATC = Air Traffic Control Operator, ISRO = Information System Radio Operator, ALS = Aircrew Life Support Specialist, PS = Personnel Specialist, PMEL = Precision Measurement Equipment Laboratory Specialist, AGE = Aerospace Ground Equipment Mechanic, p = persons, m = methods, t = tasks, i = items or steps. G Coefficient calculated for m = 2, t = 15, i = 10

One set of D study results are shown in the last line of Table 2: The generalizability of WTPT scores in the JPM system estimated by computing G coefficients for scores averaged over 2 methods, 15 tasks per method, and 10 steps per task[2]. As can be seen in the table, the G coefficients were at least .85 for six of the eight jobs, indicating that the WTPT was generally a reliable measure of job proficiency.

[2] Generalizability coefficients were calculated as the ratio of universe score variance (s_p^2) divided by total observed score variance (all variance components containing the subscript p). Technically, the discriminant validity terms (e.g., $s_{p(t:m)}^2$) could be added to observed score variance since they are hypothesized to be a type of desirable variance. We did not do so for two reasosns: first, to provide more conservative estimates of dependability; and second, to generate results which could be more comparable to traditional reliability estimates.

Job Proficiency Ratings

Generalizability Designs. To investigate the generalizability of performance ratings over measurement conditions, facets of interest were rating forms, sources, and the number of scales or dimensions nested within each form.

Rating forms comprised the first facet; task-level, dimensional, global, and Air Force-wide forms were the conditions of the facet. These can be considered random samples of a larger universe of possible forms which could be used to assess ratee performance.

The second facet was rating sources, with incumbents, peers, supervisors as the conditions of the facet. These three sources can be considered random samples of a larger universe of possible sources which could be used to assess ratee performance (e.g., Airmen could also be rated by second-level supervisors or trained observers). When Airmen were rated by more than one peer, only a single randomly-selected rating was selected to balance the design.

The final facet was the individual scales which comprised each form. The terms scales and dimensions are used synonymously to describe conditions of this facet. Scales on any one form can be considered a random sample of possible scales which could constitute a form (e.g., only a subset of all possible tasks were assessed on the task-level form). Scales were nested within forms because individual dimensions vary from form to form. The number of scales comprising a form varied considerably. Two strategies were used to balance the number of scales across forms during analyses. First, analyses were conducted with two randomly selected scales from each of the four forms (as the shortest form had only two scales). Second, the two-item global rating form was excluded and analyses were conducted using $\underline{x}$ randomly selected scales from each of the remaining three forms, where $\underline{x}$ was the number of scales on the dimensional form (the next shortest form). Results from both analyses were similar and yielded comparable conclusions regarding the generalizability of ratings. Only the results of the three-form analyses are presented in this paper since these contain less sampling error.

Results Supportive of Construct Validity. Table 1 shows the expected results for variance components which would be considered evidence of the construct-related validity of measures. The final three expected results are specific to the proficiency ratings.

G Study Results. Estimated G study variance components for the full design are presented in Table 3 for the eight occupational specialties. Again, the larger values for $\underline{s}_{p(i:t:m)}{}^2$ indicate there is at least a moderate amount of random error within the ratings in each specialty. In contrast to the results for the WTPT designs, there is greater consistency in the magnitude of variance components across specialties. Thus, it is more likely that inferences about the relative size of variance components may be generalized to additional specialties. Of interest is the main effect for rater sources, $\underline{s}_s{}^2$, a value which indicates whether there were mean differences in average ratings by different sources. This value was near zero in six specialties, though substantially larger for Air Traffic Control Operators and Personnel Specialists. This low value indicates that self ratings were not substantially higher than ratings from other sources. While there is a widely-held belief that self ratings are too lenient (e.g., Thornton, 1980), the effect size reported here is consistent with the value reported in a meta-analysis of 57 rating studies (Kraiger, 1986).

Table 3. Estimated Variance Components for G Study of Rating Variables with Three Forms.

	Job:							
	JEM	ACS	ATC	ISRO	ALS	PS	PMEL	AGE
Effect	$\underline{s}^2$	$\underline{s}^2$	$\underline{s}^2$	$\underline{s}^2$	$\underline{s}^2$	$\underline{s}^2$	$\underline{s}^2$	$\underline{s}^2$
p	.151	.120	.118	.133	.088	.047	.087	.122
s	.015	.015	.036	.001	.010	.041	.010	.016
f	.001	-.001	-.017	-.009	.001	.002	-.005	-.006
i:f	.015	.031	.040	.025	.039	.045	.049	.054
ps	.186	.173	.208	.173	.193	.172	.140	.160
pf	-.003	-.030	-.009	.021	.028	.023	.027	.022
sf	.001	-.008	.000	.003	.000	.000	-.001	.000
psf	.016	-.018	.010	.036	.061	.043	.033	.048
p(i:f)	.057	.106	.066	.089	.074	.094	.065	.055
s(i:f)	.004	.019	.000	.002	.005	.005	.002	.007
ps(i:f)	.293	.330	.285	.306	.353	.395	.322	.359
G_{1-1-8}	.381	.345	.311	.323	.208	.136	.259	.319
G_{3-4-8}	.689	.651	.611	.574	.517	.308	.585	.641

Note. JEM = Jet Engine Mechanic, ACS = Avionic Communications Specialist, ATC = Air Traffic Control Operator, ISRO = Information System Radio Operator, ALS = Aircrew Life Support Specialist, PS = Personnel Specialist, PMEL = Precision Measurement Equipment Laboratory Specialist, AGE = Aerospace Ground Equipment Mechanic, p = persons, s = sources, f = forms, i = scales or dimensions.
G_{1-1-8} calculated for s = 1, f = 1, i = 8
G_{3-4-8} calculated for s = 3, f = 4, i = 8

The $\underline{s}_p^2$ term, universe score variance, was relatively large and varied from .047 to .151. Even larger were the residual term ($\underline{s}_{ps(i:f)}^2$) and the variance component for the interaction of ratees and sources ($\underline{s}_{ps}^2$). The former term represents the person-by-source-by-items term, confounded with random error, and is uninterpretable in this design. The large value for the latter indicates that the three sources differentially ordered ratees.

As shown in Table 1, three variance components were inspected to examine construct validity. The variance component $\underline{s}_{pf}^2$ was inspected to investigate the convergent validity of the rating forms. Its value was zero in three specialties (Jet Engine Mechanic, Avionic Communication Specialist, and Air Traffic Control), and only the seventh or eighth largest variance component in the other five specialties. As a proportion of total observed score variance, this variance component ranged from .000 to .040 across specialties. Thus, consistent with the third expected outcome, ratings converged over forms.

The fourth expected result was also found. The variance component $\underline{s}_{p(i:f)}^2$ was relatively large in all specialties. It was the third largest variance component for Personnel Specialists, and the fourth largest in all others. As a proportion of observed score variance, this variance component ranged in size from .072 (for Aerospace Ground Equipment Mechanics) to .156 for (Avionic Communication Specialists). For most specialties, this ratio fell between .090 and .120, indicating that about 9 to 12% of observed score variance is due to discriminant validity across rating dimensions. In this context, the relatively large values for $\underline{s}_{p(i:f)}^2$ are desirable since they provide evidence of discriminant validity - ratees are differentially ordered by dimensions.

The final expected outcome was also obtained. The variance component for $\underline{s}_{ps}^2$ was extremely large, second in size in each specialty to the residual error term. As a proportion of observed score variance, this variance component ranged in size from .208 (for Precision Measurement Equipment Laboratory Specialists) to .307 (for Air Traffic Control Operators). Thus, approximately 20 to 30% of the variance in ratings can be attributed to differences in ratings by incumbents, peers, and supervisors. This G study outcome was supported by univariate correlations among sources. Across forms, dimensions, and jobs, the average correlation of self ratings was .24 with both peer ratings (range between .10 and .37) and supervisory ratings (range between .15 and .35), while the average correlation between peer and supervisory ratings was .31 (range between .13 and .51).

Given the strong, expected source differences, another set of analyses was performed to investigate the generalizability of ratings within rating sources. In these analyses, forms and items nested within forms were the facets of interest, and separate analyses were conducted for each rating source. After computing G study variance components, the generalizability of the proficiency ratings was estimated with one rating form of eight scales (conditions approximating typical rating scenarios in many organizations). The G coefficients ranged (across jobs) from .496 to .720 for self ratings, .670 to .796 for supervisory ratings, and .599 to .853 for peer ratings. For the latter two sources, 15 of the 16 calculated G coefficients were greater than .650. Together with the large value obtained for $\underline{s}_{ps}^2$, these within source analyses indicate that each source provided reliable ratings, even though ratings diverged across sources.

D Study Results. Generalizability coefficients were calculated for various combinations of rating conditions. Shown at the bottom of Table 3 are G coefficients for two sets of measurement conditions: A single source using a single 8-item form, and three sources using four 8-item forms (the D study which best approximates the actual measurement conditions on the JPM system).

The D study analyses revealed that measures were more reliable when ratings are averaged over multiple sources and multiple forms. With a single source using a single eight-item form, G coefficients ranged between .135 and .302. In contrast, by averaging scores over all three sources and four forms, the generalizability coefficients ranged from .388 to .641, with most values above .500. Notably, even when scores are averaged over multiple sources of error, generalizability coefficients were relatively low, indicating that only about half the variance in observed scores is due to individual differences.

IV. DISCUSSION

The present investigation used G theory to provide evidence of the reliability and construct validity of performance measures collected as part of the JPM system. Reliability evidence was generated by computing D study generalizability coefficients under specific measurement conditions of interest to the Air Force. Validity evidence was generated by forming and testing a priori expectations about specific G study variance components interpretable as evidence of convergent and discriminant validity.

Psychometric Quality of the WTPT

The results indicated that the WTPT yielded dependable proficiency scores in nearly all tested specialties. This conclusion is supported by the following results. Reliability evidence is provided by the size of the D study generalizability coefficient; under conditions approximating those used in the JPM system (two methods, 15 tasks, 10 items), G coefficients were greater than .85 in six of eight occupational specialties. That over 85% of the variance in observed scores can be attributable to individual differences in job proficiency suggests that the WTPT is a reliable method of assessing incumbent proficiency.

Construct-oriented support for the validity of the measures consists of evidence of the reliability of measures, as well as confirmation of a priori expectations of the relative size of variance components. As shown in Table 1, we predicted relatively small variance components for $\underline{s}_{pm}^2$ (convergent validity) and relatively large values for $\underline{s}_{p(t:m)}^2$ (discriminant validity across tasks).

Expectations were supported for both WTPT variance components. First, the variance component for $\underline{s}_{pm}{}^2$ was the smallest variance component in all eight specialties, indicating that WTPT scores converged across methods. Thus, it can be concluded that the interview format is a suitable surrogate for the more labor-intensive hands-on component, a finding consistent with Hedge and Teachout (1992).

Second, the variance component for $\underline{s}_{p(t:m)}{}^2$ was the second or third largest value in all eight specialties. This result supports inferences of discriminant validity, scores on one trait are not necessarily predictive of scores on another (J. S. Kane & Lawler, 1979). Thus, the WTPT is sensitive to multi-dimensional variations in performance. From a practical perspective, the pattern of relatively small values for $\underline{s}_{pm}{}^2$ and large values for $\underline{s}_{p(t:m)}{}^2$ suggest that an efficient way to construct future versions of the WTPT is to use only a single method, but maintain a high number of tasks tested within that method.

It is interesting to note that the variance component for steps (nested within tasks and methods), was a relatively large source of variance in all but two specialties. This suggests that for any particular task, steps were not of equal difficulty and could have been sampled from different universes. One explanation for this outcome is that as incumbents become more proficient at their job, certain steps on a task become automatized and thus lose their ability (relative to other steps) to discriminate between high and low performance on the task. Alternatively, just as people perform differently across tasks, they perform differently on steps within tasks, perhaps because some steps are more difficult than others, or some are more easily skipped. Note that because $\underline{s}_{p(i:t:m)}{}^2$ and $\underline{s}_{p(t:m)}{}^2$ are independent effects in the design, problems in interpreting scores at the item or step level (due to large values of $\underline{s}_{p(i:t:m)}{}^2$) does not imply scores at the task level ($\underline{s}_{p(t:m)}{}^2$) are less valid. A selection analogy would be when total test scores are valid predictors, even if response patterns on individual items are difficult to interpret.

Psychometric Quality of the Proficiency Ratings

For six of the specialties, generalizability coefficients are greater than .70 when scores are averaged over three sources, two or more forms, and eight or more scales. However, such measurement conditions represent a considerably more extensive evaluation system than would be found in most organizations. More importantly, the large variance component for the person-by-source interaction suggests that different sources provide very different ratings of incumbents and calls to question the validity of averaging over sources. If ratings are not averaged over sources, the D study generalizability coefficients are lower when randomly selecting a single source and single form (median G coefficient = .315).

On the other hand, when only a single form (with eight rating scales) is used, analyses within rater sources produced coefficients greater than .60. This value is similar to other values calculated as indices of the reliability of ratings (King, Hunter, & Schmidt, 1980) and similar in magnitude to values found in other G theory investigations of performance ratings (Day & Silverman, 1992; McHenry, Hoffman, & White, 1987; Webb, Shavelson, Kim, & Chen, 1989).

These values suggest that while ratings are not as reliable as WTPT scores, they are still somewhat reliable within sources as measures of job proficiency. While ratings within a specific source are reliable, these ratings cannot be generalized to ratings by other sources. Conceptual issues suggested by the differences between G coefficients calculated within specific sources, and those calculated when *randomly* selecting a single source are discussed below.

While proficiency ratings did not converge over sources, they did show adequate levels of discriminant validity within forms and convergent validity across forms. As expected, the value of $\underline{s}_{p(i:f)}^2$ was relatively large, the third largest value in one specialty, and the fourth largest (behind $\underline{s}_{ps(i:f)}^2$, $\underline{s}_{ps}^2$, and $\underline{s}_p^2$) in the other seven specialties. These results suggest that across forms and sources, raters are able to differentially rank-order ratees on discrete dimensions of job performance, again suggesting that the performance measures had adequate discriminant validity.

Also as expected, the value of $\underline{s}_{pf}^2$ was relatively small. Among variance components comprising observed score variance ($\underline{s}_p^2$, $\underline{s}_{ps}^2$, $\underline{s}_{pf}^2$, etc.) $\underline{s}_{pf}^2$ was the smallest variance component in all specialties. Thus, while different rating forms operationalize job proficiency differently at the scale level (e.g., specific tasks vs. global performance), measures of overall proficiency (summed over scales) converge. Together, these results suggest that: (a) given a well-designed form, raters can adequately discriminate among levels of performance in a multi-dimensional criterion space; and (b) when the issue is overall performance, different types of rating forms yield similar conclusions about ratees. Similar results were found in all tested specialties, suggesting these results may generalize to additional specialties as well.

Rating Source Effects

The greatest threat to the validity of performance ratings remains the large, predictable effect for the interaction of ratees and rating sources. Estimated variance components for the person-by-source effect indicated substantial variance due to this effect. Further, while within source analyses revealed adequate reliabilities for a particular source, G coefficients for the full design when *randomly selecting* a single source indicate that ratings do not generalize over sources.

While other studies have reported low convergence over sources (for a review, see Harris & Schaubroeck, 1988), this study provides both a direct estimate of the size of the effect, as well as its consistency across eight specialties. At the same time, we note that G theory alone is insufficient to identify the causes for these differences. We hope though that by identifying the magnitude and ubiquity of the effect, we might inspire other researchers to postulate and test hypotheses which account for these differences.

One explanation for the source effect is that ratings are a perceptual phenomenon in which raters contaminate objective observations of performance with their own perceptual biases and the demands of the particular rating system (J. P. Campbell, Dunnette, Lawler, & Weick, 1970; Guion, 1965). Alternatively, raters at different organizational levels may have differential opportunities to observe ratee performance (Zammuto, London, & Rowland, 1982) or ratees may behave differently in the company of different sources. Whatever theories are offered, it is important that they be strong enough to account for what appears to be a potent effect.

Contributions and Limitations of G Theory

Finally, this study illustrates several potential contributions and limitations of G theory as a data analysis and theory-testing tool. One contribution is that it can help decision makers to refine a measurement instrument. For example, suppose that the Air Force decided that at eight hours, the WTPT is too time-consuming for wide-spread application. By inspecting the G study variance components it can be seen that there is little variance due to the interaction of persons and methods, but more variance due to the interaction of persons and tasks. Given this data, a recommendation can be made for reducing the length of the WTPT by eliminating one evaluation method, but retaining an adequate number of tasks. A D study coefficient can be calculated to estimate test reliability under these conditions.

A limitation is that not all refinements suggested by a G theory analysis may be practically implemented. For example, D study analyses may indicate that similar reliability levels may be attained by using either two raters with three forms of two items each, or one rater with one form of 12 items. Clearly, the latter conditions would be more practical and affordable to implement in most organizations.

A second limitation, noted above, is that while a G theory investigation might tell us *that* certain measurement conditions (e.g., rater sources) affect the dependability of scores, the study may not tell us *why* those conditions matter. It still important for researchers to be well aware of the theories governing the behavior of individuals on the instruments they employ.

As illustrated above, another potential contribution is that G theory enables decision makers to assess the generalizability of a measure under conditions other than those currently in use. The Spearman-Brown prophecy formula is incapable of estimating reliability as a function of increases or decreases in the number of measurements on two or more facets at the same time. G theory extends the Spearman-Brown prophecy formula by enabling estimates of reliability while manipulating more than one condition simultaneously.

Finally, G theory forces researchers and decision-makers to explicitly address measurement issues which are too often ignored in performance measurement research. These include: What are all the conditions of measurement that could affect observations of individuals? How can these be measured and controlled? Should particular measurement conditions be considered random samples of a larger set of possible conditions or do they exhaust the set? Will the same set of conditions always be used, or might a smaller or larger set be used in the future? While G theory was used in the present context as an analysis tool, it may be equally useful during the instrument development stage when decision makers have some latitude in defining possible measurement conditions.

V. REFERENCES

Austin, J. T., Villanova, P., Kane, J. S., & Bernardin, H. J. (1991). Construct validation of performance measures: Definitional issues, development, and evaluation of indicators. Research in Personnel and Human Resources Management, 9, 159-233.

Bernardin, H. J., & Beatty, R. W. (1984). Performance appraisal: Assessing human behavior at work. Boston: Kent.

Borman, W. C. (1979). Format and training effects on rating accuracy and rating errors. Journal of Applied Psychology, 64, 410-421.

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City, IA: American College Testing Program.

Boruch, R. F., Larkin, J. D., Wolins, L., & MacKinney, A. C. (1972). Alternative methods of analysis: multitrait-multimethod data. Educational and Psychological Measurement, 30, 833-853.

Campbell, J. P. (1976). Psychometric theory. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology (pp. 122-185). Chicago: Rand-McNally.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., & Weick, K. E. (1970). Managerial behavior, performance, and effectiveness. New York: McGraw-Hill.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Crick, J. E., & Brennan, R. L. (1982). GENOVA: A generalized analysis of variance program (FORTRAN IV computer program and manual). Dorchester, Mass.: Computer Facilities, University of Massachusetts.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements. New York: Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 62, 281-302.

Day, D. V., & Silverman, S. B. (1992, August). Examining the generalizability of field performance ratings. Paper presented at the annual meeting of the American Psychological Association, Washington, D.C.

Guion, R. L. (1965). Personnel testing. New York: McGraw-Hill.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. Personnel Psychology, 41, 43-62.

Hedge, J. W., & Teachout, M. S. (1986, November). Job performance measurement: A systematic program of research and development (AFHRL-TP-86-37). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Hedge, J. W., & Teachout, M. S. (1992). An interview approach to work sample criterion measurement. Journal of Applied Psychology, 77, 453-461.

Kane, M. T. (1982). A sampling model of validity. Applied Psychological Measurement, 6, 125-160.

Kane, J. S., & Lawler, E. E. (1979). Performance appraisal effectiveness: Its assessment and determinants. In B. Staw (Ed.), Research in organizational behavior (Vol. 1). Greenwich, CN: JAI Press.

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. (1970). Issues in managerial performance: Multitrait-multimethod analyses of ratings. Psychological Bulletin, 75, 34-49.

King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in multidimensional forced choice performance evaluation scale. Journal of Applied Psychology, 65, 507-516.

Kraiger, K. (1986, April). Self, peer, and supervisory ratings of performance: So what? Paper presented at the annual meeting of the Society for Industrial/Organizational Psychology, Chicago.

Kraiger, K., & Teachout, M. S. (1990). Generalizability theory as construct-related evidence of construct validity of job performance ratings. Human Performance, 3, 19-35.

Lipscomb, M. S., & Dickinson, T. L. (1988). The Air Force domain specification and sampling plan. In M. S. Lipscomb & J. W. Hedge (Eds.), Job performance measurement: Topics in the performance measurement of Air Force enlisted personnel (AFHRL-TP-87-58). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

McHenry, J. J., Hoffman, R. G., & White, L. A. (1987, April). A generalizability analysis of peer and supervisory ratings. In G. Laabs (Chair), Applications of generalizability theory to military performance measurement. Symposium at the annual meeting of the American Educational Research Association, Washington, D.C.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 69, 147-156.

Searle, S. R. (1971). Linear models. New York: Wiley.

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

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. American Psychologist, 44, 922-932.

Stanley, J. C. (1961). Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence. Psychometrika, 26, 205-219.

Thornton, G. C. III (1980). Psychometric properties of self-appraisals of job performance. Personnel Psychology, 33, 263-272.

Webb, N. M., Shavelson, R. J., Kim, K. S., & Chen, Z. (1989). Reliability (generalizability) of job performance measurements: Navy machinist mates. Military Psychology, 1, 91-110.

Zammuto, R. F., London, M., & Rowland, K. M. (1982). Organization and rater differences in performance appraisals. Personnel Psychology, 35, 643-658.