LA-UR- 96-2256

CONF-9606256--1

Title:

Measurement Issues in Assessing Employee Performance:
A Generalizability Theory Approach

Author(s):

Blair O. Stephenson

RECEIVED

JUL 19 1996

OSTI

Submitted to:

1996 IPMAAC Conference - Boston, MA

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

Los Alamos
NATIONAL LABORATORY

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy.

Form No. 836 R5
ST 2629 10/91

MASTER

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

**Portions of this document may be illegible
in electronic image products. Images are
produced from the best available original
document.**

**Measurement Issues in Assessing Employee Performance:
A Generalizability Theory approach**

Blair O. Stephenson

Los Alamos National Laboratory

Paper presented at 1996 IPMAAC Conference - Boston, MA.

Introduction

Increasingly, organizations are assessing employee performance through the use of rating instruments employed in the context of varied data collection strategies. For example, the focus may be on obtaining multiple perspectives regarding employee performance (360° evaluation). From the standpoint of evaluating managers, upward assessments and “peer to peer” evaluations are perhaps two of the more common examples of such a multiple perspective approach.

Unfortunately, it is probably fair to say that the increased interest and use of such data collection strategies has not been accompanied by a corresponding interest in addressing both validity and reliability concerns that have traditionally been associated with other forms of employee assessment (e.g., testing, assessment centers, structured interviews). As a consequence, many organizations may be basing decisions upon information collected under less than ideal measurement conditions. To the extent that such conditions produce unreliable measurements, the process may be both dysfunctional to the organization and/or unfair to the individual(s) being evaluated. Conversely, the establishment of reliable and valid measurement processes may in itself support the utilization of results in pursuit of organizational goals and enhance the credibility of the measurement process (see McEvoy (1990), who found the acceptance of subordinate ratings to be related to perceived accuracy and fairness of the measurement process).

The present paper discusses a recent “peer to peer” evaluation conducted in our organization. The intent is to focus on the design of the study and present a Generalizability Theory (GT) approach to assessing the overall quality of the data collection strategy, along with suggestions for improving future designs.

Overview of Generalizability Theory

“Generalizability (G) theory is a statistical theory about the dependability of behavioral measurements” (Shavelson and Webb, 1991).

Specifically, GT recognizes that all measurement processes are imperfect to some degree. The approach of generalizability theory can be thought of as a hybrid combination of classical reliability theory, experimental design, and analysis of variance.

Whereas classical reliability theory (Nunnally, 1978) regards measurement error as undifferentiated (systematic error is not differentiated from random error), GT recognizes that error can stem from multiple sources and that the relative contributions to error from each source can be estimated. From a single study, error associated with each source can be estimated, as can error associated with interactions among the sources.

The essential goal of GT is to allow the decision maker to assess the dependability of a given measurement process and, if necessary, modify future designs with the goal of improving both the efficiency and dependability of the measurement process.

The process begins with the researcher conducting a G study, designed to assess the measurement qualities of a given instrument/process. Subsequently, the results of the G study are used in the context of a Decision study (D study), usually involving the collection of new data. Based on the variance components derived from a G study, various D study designs can be considered from the standpoint of obtaining a more efficient and dependable measurement design.

Data Collection

Thirty-one managers were evaluated by a subset of peers (a subset of approximately 10 other managers selected by each individual manager) on 19 items related to such issues as leadership, planning, and vision (the construct might loosely be referred to as "leadership effectiveness"). From a generalizability theory perspective, the design has two facets of generalization (items and raters) with the "leadership effectiveness" of managers as the facet of observation.¹

G Study Design

The primary purpose of a G study is to obtain variance component estimates that can be used to improve the design of subsequent Decision studies. For this reason, fully crossed designs are generally preferable, as variance components for each potential source of variance can be obtained (Cronbach, 1972).² The researcher can then estimate a variance component for all facets of generalization which could potentially be included in the design.

For example, variance components from a fully crossed G study can also be used to obtain generalizability coefficients for future studies that are not fully crossed. The downside of a fully crossed design is that from a resource and logistical standpoint, such a design is not always feasible.

In the present study, not all raters evaluated each manager. However, in many instances, raters did evaluate common managers. Accordingly, the G study for our analysis was based on a composite (average) of all possible pairs of raters who had

¹ The facet of generalization refers to any source of variation that affects the measurement of objects under study. The facet of observation refers to the object of study (in the present case, Managers). As Cardinet, Tourneur, and Allal (1976) point out, persons need not always be the object (facet) of measurement. For example, in program evaluation, classes may be the object of measurement, with persons within class treated as a measurement facet.

² An exception is when confounding can be expected in the D study. A G study with similar confounding will provide more appropriate estimates of the variance components (Cronbach, 1972).

evaluated a minimum of at least five common targets (110 such pairs).³ The resulting design allowed us to compute mean squares and estimate variance components from a 3-factor ANOVA (Manager x Rater x Item).

Results for fully crossed G Study

Table A presents the results of the G study based on the composite of 110 separate 2x(5-10)x19 ANOVAs. Recall that two raters provided ratings for a minimum of five managers on 19 items. The column labeled "Percentage of Total Variance" provides an indication of the relative importance of each facet in generating variability associated with the process of assessing leadership effectiveness. It is important to note that these variance components are with reference to a single rating on a single item by a single rater. By increasing the number of items and/or raters in our design, we can (in future studies) reduce the variance associated with raters and items (and associated interaction terms) while hopefully increasing the relative amount of variance associated with Managers.

An examination of the variance components in Table A reveals that a substantial amount of the variability associated with ratings of leadership effectiveness is due to raters (15.38%). In other words, across items and targets (Managers), raters systematically differ in their use of the scale (most are "easy" graders, but a few are "hard" graders). The percentage of variance associated with Managers (the facet of observation) is small but not insignificant (5.77%). Of course, a larger variance component would be preferable for the object of measurement. In other words, a larger variance component is indicative of differentiation among managers with respect to the construct of "leadership effectiveness".

All three 2-way interaction terms are associated with meaningful variance components, especially the Manager x Rater interaction (18.43%). The relatively high magnitude of the Manager x Rater component suggests that it will be to our advantage to include as many raters as feasible. Evidently, managers (as raters) differed in how they perceived the same managers. In many GT studies, raters are trained observers and the associated variance components are quite small (Lomax, 1982). This raises an interesting question with respect to the use of multiple raters in an employee assessment context. In many instances, raters are unlikely to receive any training in how to observe and evaluate behaviors related to job performance. Lack of training in observational techniques will likely contribute unreliability to the measurement process.

In contrast, the Manager x Item variance component is less than half of that associated with the Manager x Rater interaction (7.09). Thus, we stand to gain a more dependable measurement process by increasing the number of raters as opposed to increasing the number of items. Conversely, eliminating items will have a less adverse effect on reliability than will reducing the number of raters. The percentage of variance

³ To keep the number of conditions within the item facet consistent, missing values for individual item scores were replaced with the appropriate item average for the target (manager).

associated with items is meaningful (5.83 %). Across managers and raters, the overall scores for the 19 items exhibited a fair amount of variability. Finally, the percentage of variance associated with the Rater X Item interaction was 5.85. Thus, across all targets (Managers), there was some tendency for raters to differ in their perception of the extent to which various "leadership effectiveness" traits were present in the leadership committee as a whole. As this effect is taken across all targets, it is a constant and does not contribute to measurement error when the goal is to make distinctions among managers.

It is important to note that the variance components in Table A were not based on the entire dataset. Specifically, the fully crossed design implies that all managers were evaluated on all items by all raters. However, only a subset of raters were common to some managers. The effects of such an aggregate analysis on the accuracy and stability of the estimated variance components is not known. The limitations of our data collection strategy are apparent when we consider both the magnitude of the variance component associated with the object of measurement (Managers) and the large standard errors associated with most of the variance component estimates presented in Table A.

Manager Variance

The relative magnitude of the variance component associated with the Manager effect is smaller than one would typically desire. Ideally, we wish to maximize the variance associated with Managers ("true score" variance) while minimizing the variance associated with the facets of measurement (specifically, relative "error" variance associated with interaction terms containing the object of measurement [Manager] and facets of measurement [raters and items and random error]).⁴

To some extent, the small amount of Manager variance reflects the limited nature of our design (using only a subset of the data). However, there may well be other reasons why the variance component associated with the Manager component is so small in the present study:

- a) the population associated with the object of measurement is restricted on the trait being assessed. Generalizability coefficients, all other things being equal, will be minimized to the extent that the population of interest is homogeneous on the trait being measured (Mitchell, 1979). Presumably, one would expect

⁴ In this paper, error is conceived of as relative error (as opposed to absolute error). Managers are receiving feedback comparing their scores to their peers. At this time, no absolute interpretations of overall scores or individual item scores is being made. Within a testing context, this distinction is the basis for using absolute error for assessing the reliability of criterion-referenced tests and using relative error for assessing the reliability of norm-referenced tests. The flexibility of generalizability theory is in evidence here, for if the use of "leadership effectiveness" scores were to become criterion referenced, we would become concerned about the variance components associated with the "rater" and "item" main effects. However, if relative distinctions among managers is the goal, then main effect components represent constant effects (such as a "hard" grading rater or an "easy" item) that have no bearing on making relative distinctions among managers.

that the top 31 (out of 950 managers/supervisors with some degree of supervisory responsibility) are to some degree self-selected in terms of the construct being assessed ("leadership effectiveness").

b) this was the first year of this evaluation, managers were exposed to a new assessment process with little forewarning--there may have not been enough time for managers to develop an experiential basis for evaluating each other on some of the topics addressed. Although these managers work with each other as part of a leadership committee, they do not work with one another as typical coworkers. Thus, for some topics, managers may have had limited or "unreliable" data points upon which to base their judgments. Presumably, such a limitation would be reduced over time. Additionally, perhaps even a short discussion period in which ambiguities regarding the interpretation of specific items could be resolved would increase the variance associated with managers while simultaneously decreasing the variance associated with the measurement facets (items and raters).

c) Peers may be reluctant to critically evaluate other peers--46% of all responses were '5's, the maximum value on the 1-5 scale. Perhaps a seven point scale would enhance the potential for differentiation among managers.

Sampling Error of Variance Component Estimates

Unfortunately, the standard errors associated with our G study variance component estimates are exceedingly large.⁵ Although the total number of observations is not unreasonably small, our interest in obtaining estimates for a fully crossed design substantially limits the actual data used in estimating variance components for a fully crossed design. Using only two raters to estimate the variance associated with the rater facet is too limiting.⁶ If we project the standard errors associated with the variance components obtained with a D study design using 10 raters and 19 items, the standard errors associated with measurement facets are more reasonable (see { } std. errors in "Estimated Variance Component" column of Table A). The sampling error associated with the object of measurement (Managers) remains an issue. It is too large, relative to the magnitude of the component. Of course, as noted above, there may be features of the

⁵ Standard errors are in parentheses of "Estimated Variance Component column". Standard errors are computed from GENOVA (Crick, J. E., & Brennan, R. L., 1982). GENOVA does not accept fractional degrees of freedom, so standard errors are slightly inaccurate. Smith (1978) provides formulas for computing standard errors associated with variance component estimates.

⁶ As an alternative, a nested G study analysis was performed. Such a design uses all obtained ratings, treating raters as nested within managers. The results (not presented here) provided a similar pattern of variance component estimates with somewhat smaller standard errors than obtained from our limited sample fully crossed design. However, the nested design has the severe limitation that it does not allow the researcher to distinguish between the "Rater" main effect and the "Rater X Manager" interaction effect. It is important to be able to partition the variance associated with these terms when the focus is on making relative distinctions between managers.

data collection process itself that could be improved (using a seven point scale, providing raters with more of an experiential basis for making judgments, eliminating ambiguity with respect to specific items, selecting raters randomly rather than allowing self-selection). Factors such as these are not typically regarded as part of the "Decision" making stage of generalizability theory. However, such factors can in themselves improve the measurement process without having to explicitly alter the number of conditions within each facet (e.g. more raters and/or items).⁷

Certainly, there is nothing about this initial study that should be viewed as definitive. At best, we perhaps have some indication that a significant portion of unwanted variance is associated with the Rater x Manager interaction. Increasing the number of raters is likely to improve the dependability of the measurement process more than increasing the number of items. From a time and resource perspective, it is probably more efficient to have raters evaluate 30 managers on 10 items than to have raters evaluate 10 managers on 30 items.⁸

Decision (D) Study

A major goal of generalizability theory is to improve the efficiency of future designs while ensuring that the measurement process remains dependable. Although the variance component estimates in the present study are unstable, for expository reasons, we proceed with our D study analysis. The conclusions in the D study section should not be taken seriously, until our G study results are replicated and more stable estimates of the variance components are obtained.

Table B presents generalizability coefficients for a number of possible designs and varying levels of conditions within the facets of generalization.⁹ Design 1 refers to a design where both raters and items are conceptualized as random. In other words, the intent is to differentiate managers with respect to leadership effectiveness across all possible items (measuring leadership effectiveness) and across all possible managers who could be selected by the manager as raters.¹⁰ The row labeled Design 1 shows that in

⁷ The data collection strategy limits our capacity to include the entire dataset in assessing the reliability of the entire measurement process (raters and items as independent facets). However, if we disregard the rater facet, and assign a "score" to each target-item combination based on the average across raters, coefficient alpha is a respectable .89. Thus, the 19 items appear to be assessing the same underlying construct ("leadership effectiveness"). Of course, our G study analysis makes the same basic point: Measurement error in our study can be attributed primarily to the rater facet and not the item facet. Incidentally, coefficient alpha is algebraically equivalent to the generalizability coefficient that would be obtained from a one-facet model (items as the only measurement facet).

⁸ This assumes that there is an interest in a "leadership effectiveness" construct across all items. Given how scores are reported and utilized, this may not be a reasonable assumption in the present case.

⁹ We follow the convention of referring to agreement coefficients as generalizability coefficients when score interpretation is relative (norm-referenced). The term dependability coefficient is usually reserved for absolute score interpretation (criterion-referenced).

¹⁰ Presumably, this would reflect any possible subset of 30 colleagues who would be familiar with the "leadership effectiveness" of a given manager. The "exchangeability" interpretation of random sampling of facet conditions suggests that we could view raters as random if there exists other equally acceptable raters

order to obtain a generalizability coefficient of at least .75, fifteen raters must be utilized. In actuality, managers were rated by ten raters on 19 items (although not necessarily by the same 10 raters). Column B indicates that the generalizability coefficient is .70 for such a design.

The value of .59 in parentheses (column "B") is the generalizability coefficient when the rater facet is regarded as nested within managers. Intuitively, it makes sense that the dependability of the measurement process would decline under such a model. Remember that raters differ systematically in their use of the scale ("hard" vs. "easy" graders). When unique raters are used for each manager, the variance associated with systematic rater differences in scale use cannot be distinguished from variance associated with the result of the same set of managers being evaluated differently by different raters (Manager x Rater interaction). Nested designs (D study) are not inherently less reliable than fully crossed designs (Suen, 1990). However, the estimate of relative error variance is less precise in the nested design (it includes the variance component associated with rater main effect, along with the associated interaction term). Consequently, the generalizability coefficient will be a more conservative estimate.

How could the reliability of the measurement process in future studies (Decision studies) be improved? For instance, how would the quality of the measurement process be affected if the number of items were to be reduced (e.g. to 10)? Column E shows that with 10 raters, the G coefficient would be .66. Column F shows that with 30 raters and 10 items, the G coefficient would be .80.

Design 2 treats items as fixed and raters as random. The intent here is to differentiate managers on specific items measuring leadership effectiveness across all possible raters. As the row labeled Design 2 indicates, G coefficients are all higher for this component, although by selecting such a design, generalizing to other possible sets of "leadership effectiveness" items is not possible. Design 2 is not realistic in the present context. In most all cases, a researcher would want to generalize to other similar items assessing "leadership effectiveness".

Design 3 treats items as random and raters as fixed. For such a design, the measurement goal is to differentiate managers as they are perceived by specific raters across all possible items measuring "leadership effectiveness". This design inhibits generalization to other possible ratings by other subsets of similarly situated peers. However, this design may not be unrealistic in the present case, as managers selected their own raters and the subset of choices is not likely to change significantly in the future. In contrast to a "typical" G study, where raters are "exchangeable" as trained observers, the raters in the present study bear a unique relationship with the managers that they evaluate (i.e., they have been selected by the manager). Treating raters as fixed probably makes sense. Remember that each rater was evaluated by 10 other managers,

who could have been sampled (Shavelson & Webb, 1989). If raters (and potential raters) cannot be viewed as "exchangeable", then treating raters as a fixed facet may make sense for future design considerations.

who were selected from a potential pool of 30.¹¹ The researcher must ask the question: “Am I only interested in the opinions of these 10 managers or do I wish to generalize to all 30 potential evaluators?” Perhaps if, over time, managers are continually evaluated by the same 10 raters, then the researcher would come to regard raters as a fixed facet. If raters can be regarded as a fixed facet, then any of the designs with 10 more items have more than acceptable generalizability coefficients (all over .85).

In summary, if the decision maker believes that a measurement facet can be regarded as fixed (Design 2 or 3), measurement error will be reduced. However, one must be willing to accept the corresponding restrictions in generalizing to all items that could assess “leadership effectiveness” or all raters who could potentially evaluate managers.

Single Facet Design

The analysis thus far has focused on examining the measurement process in terms of generalizing over all items and raters for a given manager’s score. However, for the most part, managers received feedback at an individual item level. Managers never received an overall “leadership effectiveness” score.

Accordingly, it seems reasonable to examine each item individually. Within the generalizability framework, such a design would be a single-facet design, with raters as the only measurement facet. With only one measurement facet, the generalizability coefficient becomes primarily an indicator of rater reliability for each individual item taken one at a time. Such an analysis may provide an indication of which items were interpreted in different ways by different raters. A low generalizability coefficient may be an indication that either the item is ambiguous or that raters have varying degrees of an experience in dealing with other managers vis a’ vis the content area referenced by the item.

Table C presents generalizability coefficients for each of the 19 items (for scenarios based on both 10 and 15 raters). In general, the coefficients are tolerable, although there is certainly room for improvement. Even given the tentative nature of the variance components in our study, this might be a case where it would be reasonable to focus on the five most unreliable items (*) and perhaps improve the clarity of these items without incurring undue effort or cost. Perhaps these particular items are too vague, or alternatively, the items are clear but raters have had relatively fewer opportunities to observe other managers perform the various tasks assessed by certain items.

Summary

¹¹ It may be the case that not all 30 managers are capable of evaluating all other managers. Although all managers are members of the same leadership committee, the extent of their working relationship with other managers varies. To the extent that managers are not “exchangeable” as raters, the notion of treating raters as a fixed facet gains appeal.

At best, we can probably view this study as an exercise that may provide some indication of how to potentially improve an evolving assessment process. The limited sample sizes and unstable estimates are serious problems. However, at the very least, the GT approach forces us to consider how we intend to implement and interpret our assessment instrument(s).

The lack of differentiation with respect to managers, highlighted by our G study, suggests a number of easily implemented design improvements not directly related to altering the measurement facets (items or raters).

Generalizability theory offers the possibility of including multiple sources and/or instruments as measurement facets. The present study represents an initial pilot study administered to a small (and perhaps unique) subset of our workforce and one piece ("peer to peer") of our organizations movement toward implementing a 360° appraisal system for all employees. As Kraiger and Teachout (1990) point out, multiple instruments as a facet within a GT framework can in itself provide useful information regarding multiple perspectives of an employee's performance. Interestingly, these researchers found a sizable variance component associated with their Person X Source term, indicating that differing sources (self, peer, supervisors) vary in their ratings of employee performance. This disagreement, given an otherwise seemingly reliable process (i.e., consistency among raters of the same type), may be indicative of differing, but equally valid, perspectives on ratee performance and may not be true measurement error. From a GT perspective, a researcher would likely not expend future resources to reduce variance associated with sources; rather the measurement goal would be to establish that the measurement process for each source taken separately is reliable.

The rater facet raises a number of questions when applying GT to the realm of employee assessments. Unlike tests involving judges or classroom observations where raters may be viewed as interchangeable and the number of potential raters as infinite, such is not the case when it comes to employee assessment. The rater must be someone who is familiar with the job performance of the target. The number of potential raters is finite and may best be regarded as a fixed component of the measurement process. When raters are viewed as a fixed facet, rater variance is not considered error variance because the target manager's score is taken to be the average of all conceivable raters. In other words, the "chosen" raters are an inherent part of the measurement process.

Even if the rater facet can be regarded as random, it may not be feasible to supply additional raters to improve the measurement process. From a purely statistical perspective GT suggests that the present measurement process would be made more reliable by including more raters, but if each additional rater is less and less familiar with the target, we might simultaneously be working to increase rater variance at the same time that we are trying to reduce it.

Generalizability theory, unlike the classical reliability model, offers a methodology for addressing validity issues as well. As noted by Kraiger and Teachout

(1990), differing rating forms measuring the same construct could easily be treated as measurement facet(s). In this case, a small variance component associated with the Person X Forms interaction would be indicative of convergent validity. In their study on the assessment of Air Force personnel, Kraiger and Teachout found that three separate forms used to assess job performance converged (i.e., a small variance component for forms). The conditions of the forms facet represented different levels from which an evaluator could view job performance. The three forms in their study were task level (e.g., adjust engine starter pad), dimension level (e.g., inspect engine), and organization wide (e.g., technical knowledge/skill).

In conclusion, the results of the present study would have to be regarded as extremely speculative due to the large standard errors associated with the variance component estimates. However, we are still in the beginning phase of developing a measurement process for assessing employee performance from multiple perspectives. At the very least, the GT framework offers a way of conceptualizing what aspects of the measurement process are likely to be influential. In the future, we can hope to initiate a data collection strategy that will enable us to more accurately assess the relative amounts of variance associated with various measurement facets.

Table A
G Study--Fully Crossed Design

Estimated Variance Components for "Leadership Effectiveness" scale

(Results in table based on average of variance component estimates from 110 rater pairs evaluating a minimum of 5 common managers on all 19 items). G Study Std. Errors in (); D study (10 raters, 19 item) Std. Errors in { }

Source of Variation	df	Sum of Squares	Mean Squares	Estimated Variance Component	Percentage of Total Variance
Manager	5.4	25.47	4.78	.05070 (.09) {.09}	5.77
Rater	1	19.84	19.84	.13509 (.14) {.01}	15.38
Item	18	25.16	1.40	.05121 (.05) {.002}	5.83
Manager x Rater	5.4	18.36	3.44	.16187 (.10) {.016}	18.43
Manager x Item	97.2	46.37	.48	.06230 (.05) {.003}	7.09
Rater x Item	18	12.19	.68	.05141 (.04) {.0003}	5.85
Manager x Rater x Item, error	97.2	35.26	.37	.36588 (.06) {.002}	41.65

Generalizability coefficient (19 items and 10 common raters per Manager) = .70

NOTE: a general interpretation of the variance component for Manager might be as follows:

In theory, a mean score for each manager in the population could be obtained by averaging over all possible items and raters in the universe. The variance associated with this resulting distribution of mean scores is the population variance component. The variance component estimate that we actually compute from the resulting ANOVA mean squares is an estimate of this population variance component.

The associated standard deviation (.23 for Manager) is a scale based estimate of how much managers vary in "leadership effectiveness" across all items and all raters. Given a 4 point range, the expectation is that Manager scores should be distributed within a .92 point range (four std. dev* .23).

Ideally, one would hope that the variance component for the object of measurement (Manager) will be relatively large, while all other variance components (associated with facets of measurement--items and raters) will be relatively small. To the extent that variance components associated with facets of measurement are large, increasing the number of conditions within the facet (e.g., the number of raters and/or items) will reduce the relative size of the associated variance component and increase the overall accuracy of the measurement process (i.e., the generalizability coefficient).

Table B

Generalizability Coefficients for the two facet designs involving ratings of "Leadership Effectiveness" (Measurement facets are items and raters).

Design	Facets R=Rater I=Item	Object of Measurement	G Coeff.	G Coeff.	G Coeff.	G Coeff.	G Coeff.	G Coeff.	G Coeff.
			A	B	C	D	E	F	G
			1 raters/ 1 items	10 raters/ 19 items	30 raters/ 19 items	15 raters/ 19 items	10 raters/ 10 items	30 raters/ 10 items	10 raters/ 30 items
1	R-random I-random	Managers' "leadership effectiveness" whatever may be the items and raters	.08	.70 (.59)	.82	.77	.66	.80	.72
2	R-random I-fixed	Managers' "leadership effectiveness" on 19 specific items whatever may be the raters	.18	.75	.90	.82	.74	.90	.75
3	R-fixed I-random	Managers' "leadership effectiveness" as viewed by specific raters whatever may be the items	.33	.93	.93	.93	.87	.88	.95

Table C

Generalizability Coefficients for Target X Rater model
(for each item separately, coefficients can be interpreted as rater reliability)

Item	Generalizability Coefficient (10 raters)	Generalizability Coefficient (15 raters)
1) Keeps commitments as promised	.77	.83
2) Accepts responsibility for failures and successes	.71	.79
3) Actions are consistent with words	.69	.77
4) Treats others with respect	.79	.85
5) Supports and implements institutional goals and decisions	.75	.82
6) Actively works to reduce the cost of doing business	.76	.82
7) Cooperates with others to deliver product/service on time and within budget	.79	.85
8) Supports affirmative action and diversity	.82	.87
9) Models behavior expected in others	.64 (*)	.73
10) Keeps issues above the table (open communication with no "end runs")	.64 (*)	.72
11) Willingly collaborates with others	.71	.78
12) Takes a Laboratory-wide viewpoint on institutional issues	.78	.84
13) Considers alternative points of view in decision making	.69	.77
14) Is consistently effective as a team player	.69	.77
15) Has professional expertise to contribute to LLC decisions	.62 (*)	.71
16) Contributes to reaching solutions	.76	.83
17) Accepts others' ideas and builds on them	.66 (*)	.75
18) Seeks feedback to improve individual performance	.80	.86
19) Supports in word and action the Laboratory mission, vision, and values	.74	.81

References

- Brennan, R. L. (1983). Elements of generalizability theory. Iowa City, IA: American College Testing Program.
- Cardinet, J., & Tourneur, Y. (1976). The symmetry of generalizability theory: Applications to educational measurement. Journal of Educational Measurement, 13, 119-135.
- Crick, J. E., & Brennan, R. L. (1982). GENOVA: A generalized analysis of variance system (Fortran IV computer program and manual.) Dorchester, MA: University of Massachusetts at Boston, Computer Facilities.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.
- McEvoy, G. M. (1990). Public sector Managers' Reactions to Appraisals by Subordinates. Public Personnel Management, 19, 201-212.
- Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychological Bulletin, 86, 376-390.
- Nunnally, J. C. (1978). Psychometric Theory. New York: McGraw-Hill.
- Shavelson, R.J., Webb, N.M., & Rowley, G.L. (1989). Generalizability Theory. American Psychologist, 44, 922-932.
- Smith, P. L. (1978). Sampling Errors of Variance Components in Small Sample Multifacet Generalizability Studies. Journal of Educational Statistics, 3, 319-346.