DOCUMENT RESUME

ED 301 818 CG 021 306

AUTHOR Wilson, F. Robert; Yager, Geoffrey G.
TITLE Generalizability of Effectiveness Ratings for Counselors and Their Supervisors.
PUB DATE Oct 88
NOTE 28p.; Paper presented at the Annual Meeting of the National Association for Counselor Education and Supervision (St. Louis, MO, October 7-10, 1988).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Counselors; *Evaluation Methods; *Rating Scales; *Supervision; *Supervisory Methods

ABSTRACT

It is common for supervisors to evaluate their supervisees with a rating form. Despite the importance of supervisor ratings to the training of counselors and therapists, very little attention has been devoted to the overall reliability (generalizability) of these ratings. This study examined the generalizability of supervisor ratings of counselors-in-training. Participants included 23 counselor trainees enrolled in a masters level prepracticum course and 9 doctoral-level counseling supervisors. Ratings of counselor and supervisor effectiveness were collected through the use of the Counselor Effectiveness Scale. At the beginning of the term, practicum trainees were randomly assigned to supervisors. Each prepracticum counselor audiotaped a counseling session with a volunteer client on each of 6 weeks. Within a week following each counseling session, counselors met with their supervisors for a 60 minute supervision session. Following each supervision session, the supervisor rated the effectiveness of the counselor, and the counselor rated the effectiveness of the supervisor. Generalizability analyses were performed. Results showed generalizability of supervisor's ratings of counselor effectiveness were affected more by the number of occasions on which the counselor was rated than by the length of the rating instrument. Similar findings were observed for counselor ratings of supervisor effectiveness. (ABL)

Generalizability of Effectiveness Ratings for Counselors and Their Supervisors

F. Robert Wilson and Geoffrey G. Yager

University of Cincinnati

RUNNING HEAD: Generalizability of Effectiveness Ratings

A Paper Presented at the First National Convention of the Association for Counselor Education and Supervision
St. Louis, Missouri
October, 1988

Abstract

The generalizability of counselor and supervisor effectiveness ratings was investigated through data collected during supervision activities within a counseling practicum class. Generalizability of supervisor's ratings of counselor effectiveness were affected more by the number of occasions on which the counselor was rated than by the length of the rating instrument. Similar findings were observed for counselor ratings of supervisor effectiveness.

Generalizability of Effectiveness Ratings

for Counselors and Their Supervisors

It is common, if not required, for supervisors to evaluate their supervisees with a rating form. As Stoltenberg and Delworth (1987) have recently indicated, it is such quantitative evaluation of a counselor's abilities that may comprise the full extent of information available to a new supervisor in planning the continued training of that counselor. "Although it is a common assumption that one can sort a room of therapists into good ones and others, the danger exists that the interpretations or inferences made by a supervisor may be misleading" (Stoltenberg & Delworth, 1987, p. 113).

Despite the importance of supervisor ratings to the training of counselors and therapists, very little attention has been devoted to the overall reliability (generalizability) of these ratings. The primary objective of this investigation was to study the generalizability of supervisor ratings of counselors-in-training. Specifically, the study was designed to answer the following basic question: If we were to devise a optimally effective rating scale for a supervisee effectiveness rating, how many items would be rated on that scale, and how often (i.e., on how many different occasions) would we ask the supervisor to make those ratings?

A second type of rating form common in supervision research is a trainee assessment of the supervisor. Stoltenberg and Delworth (1987) argue that trainee ratings may well be much more variable than those of supervisors since they are often a

function of "how comfortable--and not how effective--supervision was" (p. 118). They suggest that some considerable time may pass after supervision is completed before some trainees can best recognized the impact that supervision has had upon their learning. Thus, it appeared very relevant to include in this investigation a second important research question: What is the optimal number of times, occasions, and counselor/supervisees to discriminate among supervisors of varying abilities and characteristics?

To address each of these questions, generalizability analysis was employed. Generalizability theory liberalizes and extends classical test theory. In particular, it allows for consideration of multiple sources of error by applying analysis of variance procedures to assess the dependability of measurements (Brennan, 1983; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Webb, Rowley, & Shavelson, 1988). Consequently, generalizability theory is applicable to a broad range of measurement, evaluation, and testing studies that arise in education and psychology. The counseling supervision literature, to this point, included no generalizability analyses of ratings of supervisee or supervisor effectiveness.

Generalizability research begins by conducting a generalizability study (G-study). A design for data collection is constructed to include the facets (i.e., variables) of interest. For example, a researcher might construct a design to

investigate the generalizability of trained raters' evaluations of counselor interventions. If the researcher were interested in determining the generalizability of ratings across multiple raters and multiple constructs (e.g., empathy, genuineness, regard, concreteness), the design would include a two measurment facets for raters and constructs, and a single differentiation facet, counselors. Data generation might involve having multiple raters rateeach of anumber of counselor's taped interviews on each of the constructs of interest.

On completion of the G-study, generalizability analysis allows the estimation of generalizability coefficients for hypothetical designs (D-studies) using the variance components computed from observed data (the G-study). This use of generalizability analysis is analogous to classical measurement theory's use of the Spearman-Brown prophesy formula to estimate the change in the reliability of a test if the number of items is increased or decreased. However, whereas classical methods are unidimensional (e.g., items are viewed as a single source of error), generalizability analysis permits multidimensional assessment of factors affecting one's ability to differentiate reliably among the objects of measurement (Brennan, 1983; Webb, Rowley, & Shavelson, 1988).

Methods

Participants

The initial participants in this study included 23 counselor

trainees (4 male, 19 female) enrolled in a masters level prepracticum course and 9 doctoral-level counseling supervisors (4 male, 5 female). Most prepracticum students werr enrolled in their first term of work in the master's degree program in counseling. About half of the 23 students had previously held a position that was counseling-related to some extent. The doctoral level supervisors were currently enrolled in a course in counselor supervision. Trainees were randomly assigned to supervisors at the start of the practicum course.

Instruments

Ratings of counselor and supervisor effectiveness were collected through use of the Counselor Effectiveness Scale (CES, Ivey & Authier, 1978). This instrument consists of two forms with 25 semantic differential scaled items on each form. The first set of CES items, Form A, was used by supervisors to rate their supervisee's effectiveness, while the second set, Form B, was used by the supervisees to evaluate the effectiveness of their supervisors. In a study of the concurrent validity of counselor effectiveness instruments, Wilson and Yager (1987) found the two CES scales to be highly correlated (r = .94, p < .001). A more extensive review of the measurement characteristics of this instrument has been presented by Ponterotto and Furlong (1985).

Procedures

At the beginning of the term, practicum trainees were

randomly assigned to supervisors. Each prepracticum counselor audiotaped a counseling session with a volunteer client on each of six weeks. Within a week following each counseling session, counselors met with their supervisors for a 60 minute supervision session. Following each supervision session, the supervisor rated the effectiveness of the counselor (CES - Form A) and, in turn, the counselor rated the effectiveness of the supervisor (CES - Form B). These ratings were returned directly to the researcher to insure confidentiality. During the course of the investigation, some of the rating forms were not obtained after every supervisory session. As a result, not all supervisory sessions were rated, and the final number of trainees who had been rated at least six times was only 21 of the original 23 students.

To address the generalizability of ratings of counselor and supervisor effectiveness, estimated mean squares, variance components, error variances, and generalizability coefficients were computed using the General Purpose Analysis of Variance System, Version 2.2 (GENOVA, Crick & Brennan, 1984). Additional exploration of the reliabilities, correlations among variables, and mean differences among supervisors was accomplished through the use of appropriate SPSS (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975) subprograms.

In conducting the generalizability analysis, the item facet was a random facet since the rating forms used in this study were

constructed by randomly sampling items from the original CES item set. The facets, counselor, supervisor, and occasion, were treated as random facets in that counselors, supervisors, and occasions not observed in the present G study could be exchanged with those observed in the G study, even though not sampled randomly (Webb, Rowley, & Shavelson, 1988).

Results

Supervisor's Ratings of Counselor Effectiveness

To assess the generalizability of supervisor's ratings of counselor effectiveness, the data were cast in a design with one object of measurement, counselors, and two facets, occasions and items, in the design over measures. Although counselors were nested within supervisors, GENOVA (Crick & Brennan, 1984) does not permit an object of measurement to be nested within another factor, thus supervisors was not included as a facet in this analysis.

The mean squares, and estimated mean square variance components are presented in Table 1. In this G study, the predominant sources of variance are those involving counselor as a main effect (C: 30.74%) or as a member of an interaction term (CO: 13.96%, CI: 7.40%, and COI: 40.56%). These four sources of variance, taken together, account for 92.66% of the total variance. Occasions (O: 3.54%), items (I: 3.64%), and their interaction (OI: 0.16%), when taken as a set, accounted for 7.34% of the variance.

The major question to be addressed through D study analyses was whether greater generalizability could be obtained by increases in the number of items given in a single administration (the classical strategy for improving the measurement of counselor effectiveness) or by augmenting the number of occasions on which the items are administered. In addition, the question of whether greater benefit is obtained by using randomly sampled items (items nested in observations) rather than a fixed sample of items (items crossed with observations) was of interest.

Effect of Changes in the Number of Items and Occasions. To study the effect of changes in the number of items and occasions, an arbitrary minimum case of four items administered on two occasions was selected. Items and occasions were then successively doubled until an arbitrary maximum case of 64 items administered on 8 occasions was reached. Generalizability coefficients for designs featuring various combinations of sample sizes for items crossed with occasions are presented in the first half of Table 2. Increases in either the number of items or in the number of occasions produce increases in the generalizability coefficient. However, the generalizability increases more rapidly with increases in the number of occasions than with increases in the number of items. Assuming .90 as a minimally acceptable value for the generalizability of supervisor ratings of counselors, one would collect ratings on at least 8 occasions using an instrument consisting of at least 8 items.

Fixed vs. Randomly Sampled Items. Most counselor effectiveness studies use a single rating form consisting of a fixed list of items. To explore the effect of constructing a large item pool and randomly sampling a subset of items for each administration of the rating form, the data were recast into a design featuring items nested in occasions. Generalizability coefficients for various combinations of sample sizes for items nested in occasions are presented in the second half of Table 2. Element by element comparison of the crossed vs. nested designs revealed that nesting items in occasions produced a uniform decrease in generalizability over that observed for the corresponding case using a fixed item set.

Counselor's Ratings of Supervisor Effectiveness

To assess the generalizability of counselors' ratings of supervisor effectiveness, the data were cast in a design with one object of measurement, supervisors, and two facets, occasions and items in the design over measures. In this analysis of the discriminations among supervisors, the factor, counselors, was considered to be a measurement source nested within supervisors. Since not all of the 9 supervisors had complete data sets (a complete set would consist of counselor ratings from 3 different counselors -- 25 items over 6 occasions), one supervisor was dropped and ratings from five counselors were discarded, leaving a data set of 8 supervisors with counselor ratings from 2 counselors consisting of 25 items over 6 occasions. Because no

appreciable difference was found with the counselor effectiveness data in comparing items crossed with occasions versus items nested in occasions, only the analysis of items crossed with occasions was conducted with the supervisor effectiveness data.

Effect of Changes in the Number of Items, Occasions, and Counselor/Supervisees. Since the analysis of supervisor ratings of their supervisees revealed that a range of from 8 to 32 items was a sufficient spread to understand the effect of item length, an arbitrary minimum case of 8 items administered on 2 occasions by 2 counselor/supervisees was selected. Items, occasions, and counselor/supervisees were then successively doubled until an arbitrary maximum case of 32 items administered on 16 occasions by 16 counselor/supervisees was reached.

The mean squares, and estimated mean square variance components are presented in Table 3. As before, the predominant sources of variance are those involving counselor as a main effect (C:S, 10.98%) or in interaction with instrument facets (CO:S, 19.10%, CI:S, 13.55%, and COI:S, 45.99%). These four sources of variance, taken together, account for 89.62% of the variance. Surprisingly, factors involving supervisor as a main effect (S: 1.66%) and in interaction with instrument facets (SO, 0.88%, SI, 0.40%, and SOI, 1.85%) only accounted for an aggregate of 4.79% of the variance. Occasions (O, 0.00%), items (I, 5.32%), and their interaction (OI, 0.27%), taken as a set, accounted for 5.60% of the variance.

Generalizability coefficients for designs featuring various combinations of sample sizes for counselors (with various numbers of observations nested within counselors) and for items are presented in Table 4. Increases in either the number of items, the number of occasions, or the number of counselor/supervisees produce increases in the generalizability coefficient. The generalizability increases least rapidly with increases in the number of items and most rapidly with increases in the number of counselors. With ratings from 16 counselors per supervisor on at least 16 occasions using an instrument consisting of 32 items, only marginal generalizability ($G = 0.66$) is achieved. The pattern of change in this matrix of coefficients suggests that greater generalizability would be achieved by further increases in the number of counselors providing ratings, and secondarily by further increases in the number of occasions on which ratings are collected. The classical strategy of adding items to the rating instrument would clearly not be supported by the data in this case.

Classical Instrument Performance Indices

To permit relating the findings of this study directly to traditional work in the field of counselor and supervisor effectiveness, several analyses were performed based on classical true-score test theory. Item homogeneities were computed by Cronbach's α, Pearson product-moment correlations were computed among ratings made of supervisee and supervisor effectiveness

across occasions, and tests were made to determine whether there were mean differences between supervisor ratings of their supervisees and between ratings received by supervisors from their supervisees.

Item Homogeneities. Cronbach's α reliabilities and intercorrelations among variables were computed and cast as a multitrait-multimethod matrix as presented in Table 5. At each occasion, the supervisor's ratings of the counselor and the counselor's ratings of the supervisor yielded remarkably high scale homogeneities (median: .95). This finding is consistent with previous research on the Ivey scales (Wilson & Yager, 1987).

Correlations among Ratings. Supervisor ratings of counselors across occasions were all highly correlated (all were significant at $p < .01$, half were significant at $p < .001$). Their values ranged from .56 to .80 with a median of .65. These correlations tended to follow a pattern: ratings made during adjacent time periods tended to be more highly correlated with each other than were ratings made at periods more widely separated in time. Thus, although the supervisor's view of the counselor was relatively consistent from one time period to the next (median correlation between adjacent time periods: .72), there was a gradual change over time such that the supervisor's initial rating accounted for 36% of the variance in the supervisor's final rating of the counselor.

Counselor ratings of supervisors were less well correlated

(only half were significant at $p < .05$). Their values ranged from .05 to .89 with a median of .37. These correlations were uniformly patterned such that the magnitude of the correlation decreased with increases in temporal distance. The counselor's view of the supervisor was also relatively consistent from one time period to the next (median correlation between adjacent time periods: .74), but there was much more change in view over time. The counselor's initial rating only accounted for 2.5% of the variance in the counselor's final rating of the supervisor.

There was little relationship between the supervisor's ratings of the counselor and the counselor's ratings of the supervisor. These multi-rater/multi-occasion correlations ranged from -.31 to .35 with a median of .08. None were significant at $p < .05$. No clear pattern emerged among the correlations. These correlations suggest that there was no systematic mutuality among supervisor's and counselor's ratings of one another.

Differences Among Supervisor's Ratings of their Counselors. To determine whether supervisors differed, the average rating given to their supervisees at each occasion was calculated. Six analyses of variance were computed, each featuring one factor in the design over subjects, supervisors. The results of this analysis are presented in Table 5. Initially, supervisors differed in the mean effectiveness rating given to their set of supervisees, however, over the six occasions, this difference

diminished such that by the sixth occasion, no significant difference was found. Inspecting the grand mean across occasions, it is interesting to note that on each successive occasion, the overall evaluation of the entire set of supervisees became more positive (reflected by a steadily decreasing score).

Differences Among Counselor's Ratings of their Supervisors. A different picture emerged when mean ratings of supervisor effectiveness were compared across supervisors. To determine whether supervisors differed the average rating received from their supervisees at each occasion, six analyses of variance were computed, each featuring one factor in the design over subjects, supervisors. The results of this analysis are also presented in Table 5. No significant difference between supervisors was observed for any occasion when mean supervisee ratings of the supervisors' effectiveness was compared. However, inspection of the grand mean reveals that in general, there was slight but steady improvement over time (as reflected by decreasing scores) in the supervisee's perception of the supervisor.

Discussion

In response to the initial questions raised in this study, it appears relatively clear that rating scales similar to those that supervisors commonly employ to rate the effectiveness of counselor trainees can be developed to allow for good generalizability. Over 90% of the variance in supervisors' ratings of counselors is attributable to differences among

counselors (c, 30.77%) and interactions of counselors with occasions and/or items (CO, 13.96%; CI, 7.40%; COI, 40.56%). Increasing the number of occasions for observation and rating of a trainee is, apparently, more important that increasing the number of items in the rating scale. In fact, an 8 item scale administered over eight occasions is only slightly less reliable than a 64 item scale administered over the same number of occasions.

The proportion of variance accounted for by differences among counselors nested within supervisors (C:S, 10.98%) was considerably greater than that accounted for by differences among the supervisors (S, 1.66%). Furthermore, the interactions between occasions and/or items with counselors (CO:S, 19.10%; CI:S, 13.55%; COI:S, 45.99%) accounted for considerably larger proportions of the variance than does the corresponding interactions with supervisors (SO, 0.88%; SI, 0.40%; SOI, 1.85%). These findings are consistent with Stoltenberg and Delworth's (1987) speculations that trainee ratings may well be much more variable than those of supervisors since they are often a function of comfort with supervision -- not with it's effectiveness. It certainly seems that the counselor-in-training's ratings of the supervisor are much less generalizable.

Additionally, there is little correlation among the counselors' ratings of supervisor effectiveness across time. The counselors generally view their supervisors as very competent

(overall means are consistently positive), however their reasons for rating the supervisor appear to differ from time to time (very low correlations between ratings across time).

In evaluating the effectiveness of supervisors, especially heterogeneous samples of supervisors (e.g., supervisor trainees), contrary to what one might expect from classical theory, lengthening the test does not appear to be the best way to improve the dependability of discriminations. Greater dependability derives from using a greater number of counselors per supervisor and collecting ratings on a greater number of occasions. Although this approach is most desirable from a statistical point of view, it is, unfortunately, more expensive in terms of time and effort on the part of the supervisors and counselors. However, failure to take these sources of error into account would run the risk of making decisions about relative effectiveness of supervisors based on error-prone data.

Limitations

This study of supervisor evaluations of prepracticum counselors and their evaluations of their supervisor's was conducted within a single counselor training program under actual counselor training conditions. The findings may not apply (a) to training programs draw from a different population of trainees, (b) to counselor-trainees at higher levels of training, (c) to faculty rather than student supervisors, or (d) to counselors and supervisors in clinical rather than training settings.

Findings relating to the generalizability of counselor evaluations of supervisor effectiveness may be spuriously conservative due to a restriction of range in supervisor ability. Unlike field conditions, these data were collected in a training institution using supervisors who were undergoing training in supervision and were, themselves, being supervised. Each supervisor had a clear idea of what was to be the focus of the student's learning in the prepracticum class, and, it is likely, that each supervisor approached the student supervision sessions with similar goals and objectives. Under such conditions, it is reasonable to assume that their performances as supervisors would be more similar than would the performances of randomly sampled field supervisors. Under ideal circumstances, study of the generalizability of supervisor effectiveness ratings would be conducted with a heterogeneous, rather than homogeneous, sample of supervisors. Since supervisor homogeneity increases the difficulty of reliable differentiation among supervisors, it is likely that in field situations, with a more heterogeneous sample of supervisors, one would not need as many counselors and occasions as were indicated in these data.

This same homogeneity of supervisors also provides a possible explanation for the relatively small intercorrelations between the repeated ratings of the supervisor by the counselors. The larger the variability of the supervisors (as targets of the rating process), the larger the expected reliabilities. A

restricted range may, thereby, have depressed the possible intercorrelations across occasions.

Cautions

Like prophesies made through use of the Spearman Brown formula, the prophesies made in generalizability analysis relate to conditions not actually sampled (D study results). Therefore, these data require replication to determine whether the benefits anticipated will be realized in subsequent research studies. This caution especially true when the original G study was based on small samples for some or all factors under investigation, as was the case in the present study. Ideal G studies involve large samples over all facets. In this study conducted within a single counselor training program under realistic training conditions only 9 supervisors and 23 counselors-in-training were available for study within a single year.

Suggestions for Future Research

There are a number of additional research studies that would follow directly from this investigation. Among the possible next steps would be the following:

1. Add a source facet (supervisor rating, self-rating, client rating, observer rating) to the design to permit study of the facets: counselors x sources x occasions x items.

2. Increase the G study sample sizes (e.g., at least 3 or 4 counselors per supervisor, at least 6 observations, and 10

or more supervisors).

3. Use a smaller instrument (e.g., a random sample of 10 items from each of Ivey's two forms pooled into a single form, or, a simple random sample of 15 or 20 items)

4. Collaborate with other counselor training institutions to aggregate greater numbers of supervisors and counselors.

5. Collaborate with community agencies to study a heterogeneous group of supervisors employing a design featuring facets for supervisor x counselor:supervisor x occasions x items.

This study has presented perhaps the first application of generalizability analysis to the ratings of counselors and supervisors. As the first investigation in this area, it leaves a number of very interesting possible directions of research for the future.

References

Brennan, R. L. (1983). Elements of Generalizability Theory. Iowa City, Iowa: ACT Publications.

Crick, & Brennan, R. L. (1984). GENOVA: General Purpose Analysis of Variance System, Version 2.2. Dorchester, Mass.: Computer Facilities, University of Massachusetts at Boston.

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

Ivey, A. E., & Authier, J. (1978). Microcounseling: Innovations in interviewing, counseling psychotherapy, and psychoeducation (2nd Ed.). Springfield, IL: Charles C. Thomas.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. (1975). Statistical Package for the Social Sciences (2nd ed.). New York: McGraw-Hill.

Stoltenberg, C. D., & Delworth, U. (1987). Supervising counselors and therapists: A developmental approach. San Francisco, Ca: Jossey-Bass.

Webb, N. M., Rowley, G. L., & Shavelson, R. J. (1988). Using generalizability theory in counseling and development. Measurement and Evaluation in Counseling and Development, 21, 81-90.

Wilson, F. R., & Yager, G. G. (1987, November). Evaluations of videotaped counselors on a variety of counselor assessment scales. Paper presented at the Annual Meeting of the North Central Association for Counselor Education and Supervison, St. Louis, MO. (ERIC Document Reproduction Service No. ED 290 986)

Table 1

Expected Mean Squares for Supervisor's Ratings of Counselor Effectiveness (G Study: Counselors x Occasions x Items Design) (n_C = 21 Counselors, n_O = 6 Occasions, n_I = 25 Items).

Sources	df	SS	MS	EMS	%EMS
C	20	1505.99	75.2997	.4588	30.74%
O	5	168.10	33.6194	.0529	3.54%
I	24	195.97	8.1653	.0544	3.64%
CO	100	581.2965	5.8130	.2083	13.96%
CI	480	608.67	1.2681	.1104	7.40%
OI	120	78.54603	.6546	.0023	.1[illegible]%
COI	2400	1452.89	.6054	.6054	40.56%
Total	3149	4591.468	---	---	

Table 2

Generalizability Coefficients for Supervisor's Ratings of Counselor Effectiveness (n_C = 21 Counselors).

D Study: Counselors x Occasions x Items Design

Occasions	Items 4	8	16	32	64
2	.69	.75	.78	.80	.81
4	.80	.84	.87	.88	.89
8	.86	.90	.92	.94	.94

D Study: Counselors x Occasions x Items:Occasions Design

Occasions	Items 4	8	16	32	64
2	.67	.72	.75	.76	.77
4	.80	.84	.86	.87	.87
8	.89	.91	.92	.93	.93

Table 3

Expected Mean Squares for Counselor's Ratings of Supervisors' Effectiveness (G Study: Supervisors x Counselors:Supervisors x Occasions x Items Design) (n_S = 8 Supervisors, n_C = 2 Counselors per Supervisor, n_O = 6 Occasions, n_I = 25 Items).

Sources	df	SS	MS	EMS	%EMS
S	7	257.24	36.7490	0.0218	1.66%
C:S	8	236.22	29.5271	0.1440	10.98%
O	5	36.02	7.2044	(0.0)	
I	24	204.86	8.5360	0.0698	5.32%
SO	35	262.27	7.4936	0.0116	0.88%
SI	168	299.12	1.7805	0.0052	0.40%
CO:S	40	274.64	6.8661	0.2505	19.10%
CI:S	192	320.53	1.6694	0.1777	13.55%
OI	120	85.03	0.7086	0.0036	0.27%
SOI	840	547.42	0.6517	0.0242	1.85%
COI:S	960	579.11	0.6032	0.6032	45.99%
Total	2399	3102.48			

Table 4

Generalizability Coefficients for Counselor's Ratings of Supervisor Effectiveness (n_S = 8 Supervisors, n_C = 2 Counselors per Supervisor, n_O = 6 Occasions, n_I = 25 Items).

D Study: Supervisors x Counselors:Supervisors x Occasions x Items Design

Coun-selors	Occas-ions	Items 8	16	32
2	4	0.14	0.16	0.16
2	8	0.17	0.18	0.19
2	16	0.19	0.20	0.21
4	4	0.25	0.26	0.27
4	8	0.29	0.30	0.31
4	16	0.31	0.33	0.34
8	4	0.38	0.40	0.42
8	8	0.43	0.46	0.47
8	16	0.47	0.49	0.50
16	4	0.52	0.55	0.56
16	8	0.59	0.61	0.62
16	16	0.62	0.64	0.66

Table 5

Scale Homogeneities, Correlations between Ratings, and Tests for Differences among Supervisors for (a) Client's Ratings of Counselor Effectiveness, (b) Supervisor's Ratings of Counselor Effectiveness, and (c) Counselor's Ratings of Supervisor Effectiveness (n = 21 Counselors).

	S>C[1] 1	S>C 2	S>C 3	S>C 4	S>C 5	S>C 6	C>S[2] 1	C>S 2	C>S 3	C>S 4	C>S 5	C>S 6
S>C 1	(.96)											
S>C 2	.72	(.96)										
S>C 3	.70	.65	(.95)									
S>C 4	.62	.57	.58	(.97)								
S>C 5	.74	.64	.72	.80	(.96)							
S>C 6	.57	.68	.56	.64	.78	(.95)						
C>S 1	.32	.17	-.08	-.08	.07	.08	(.93)					
C>S 2	.25	.18	.09	-.31	-.19	-.18	.59	(.91)				
C>S 3	.21	.15	.04	.05	.35	.30	.28	.74	(.95)			
C>S 4	.15	.08	-.10	-.03	.17	.22	.29	.54	.53	(.94)		
C>S 5	.24	- .08	-.18	.06	.21	.20	.20	.37	.36	.89	(.96)	
C>S 6	.12	- .17	-.30	-.07	-.06	-.18	.05	.17	.12	.64	.76	(.95)
Grand M	2.99	2.71	2.61	2.55	2.26	2.23	1.90	1.90	1.81	1.78	1.67	1.60
Grand s	0.82	0.86	0.81	0.94	0.78	0.86	0.60	0.64	0.72	0.73	0.65	0.64
F(8,14)[3]	6.39	9.11	3.90	2.65	3.46	1.74	0.83	1.06	1.18	1.25	1.49	0.80
p	.001	.001	.01	.05	.02	.19	.59	.44	.38	.34	.25	.61

[1] S>C: Supervisor (S) ratings of their Counselor Supervisees (C) over six counseling/supervision sessions.
[2] C>S: Counselor (C) ratings of their Supervisor (S) over six counseling/supervision sessions.
[3] Tests for differences among supervisors mean ratings of their counselor/supervisees and of their mean rating received from the counselor supervisees.

Critical Values: $r = .43$, $p < .05$; $r = .56$, $p < .01$; $r = .66$, $p < .001$.

Note: Entries on principal diagonal (in parentheses) are homogeneity estimates.