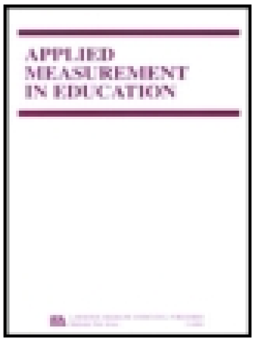

APPLIED MEASUREMENT IN EDUCATION

Applied Measurement in Education

ISSN: 0895-7347 (Print) 1532-4818 (Online) Journal homepage: https://www.tandfonline.com/loi/hame20

Models of Generalizability Theory in Analyzing Existing Faculty Evaluation Data

Lei Chang & Dennis Hocevar

To cite this article: Lei Chang & Dennis Hocevar (2000) Models of Generalizability Theory in Analyzing Existing Faculty Evaluation Data, Applied Measurement in Education, 13:3, 255-275, DOI: 10.1207/S15324818AME1303_3

To link to this article: https://doi.org/10.1207/S15324818AME1303_3

Published online: 07 Jun 2010.

APPLIED MEASUREMENT IN EDUCATION, *13*(3), 255–275

Models of Generalizability Theory in Analyzing Existing Faculty Evaluation Data

Lei Chang
Department of Educational Psychology
Chinese University of Hong Kong

Dennis Hocevar
Department of Educational Psychology
University of Southern California

In this article, we demonstrate the use of generalizability theory in analyzing existing faculty evaluation data. Three measurement conceptualizations representing different purposes of faculty evaluation were conceived to account for the existing data structure. Variance components associated with these conceptualizations were estimated from random samples taken from an existing faculty evaluation conducted in a university in the western United States. Within each of the 3 conceptualizations, 3 decision study considerations are presented together with the generalizability coefficient estimates. Practical implications for faculty evaluations are discussed.

The purpose of this article is to present three models of generalizability (G) theory in analyzing existing faculty evaluation data. These models of G theory are presented as modules to be readily applied to existing faculty evaluation, which, in most universities, is regularly conducted but not necessarily as a G study. G theory has been extensively dealt with by Brennan (1983, 1992) and Shavelson and Webb (1988, 1991), in addition to its more comprehensive treatment by Cronbach, Gleser, Nanda, and Rajaratnam (1972). The cognitive framework, notations, and symbolism of G theory used in this article are adopted from Brennan. Four previous studies (Gillmore, Kane, & Naccarato, 1978; Kane & Brennan, 1977; Kane, Gillmore, & Crooks, 1976; Smith, 1979) that applied G theory to faculty evaluation

Requests for reprints should be sent to Lei Chang, Department of Educational Psychology, Chinese University of Hong Kong, Shatin, N. T., Hong Kong. E-mail: leichang@cuhk.edu.hk

have also laid the foundations for much of the work presented here. The remainder of the article is organized according to the distinction in G theory between generalizability and decision (D) studies. In the G study, we present three conceptualizations of faculty evaluation as well as estimated variance components associated with these conceptualizations. In the D-study section, different decision considerations regarding each of the G-study conceptualizations are discussed. G coefficients associated with these D-study considerations are also provided.

G STUDY

The first task in a G study is to conceptualize the measurement under consideration by defining the objects of measurement and the conditions that are acceptable to the researcher as ways to take observations of the objects of measurement. The set of these conditions constitutes the universe of admissible observations. The same framework applies to existing faculty evaluation except that the measurement conceptualization has to be reformulated to account for the existing data structure. That is, to use the existing faculty evaluation, a researcher has to accept all or some of the measurement conditions by which the evaluation has been obtained. These conditions are discussed in the following section.

Measurement Conditions of Existing Faculty Evaluation

In a typical faculty evaluation, a faculty member is evaluated for the courses she or he has taught by the students enrolled in these courses using a standard set of evaluation items. The same evaluation procedure using the same set of items is usually repeated over semesters. As a result, there are multiple evaluations of a teacher teaching the same or different courses over time. There are also multiple evaluations of a particular course taught by the same or different teachers over time. Thus, there are five levels by which the evaluation data can be categorized. They are the teachers (t), courses (c), students (s), items (i), and occasions (o). For a given evaluation, different students are usually enrolled in different courses taught by different teachers. Thus, students are nested within teachers and within courses. For some data where different teachers teach the same courses, teachers are nested in courses. If the courses are repeated over the years, they are crossed with the occasions, whereas teachers are nested within occasions because different rather than the same teachers teach the courses over the semesters. Students who are nested in teachers are also nested in occasions.

In other data, when a teacher teaches different courses, courses are nested in teachers. Because the same teachers teach different courses over the semesters, the teachers are crossed with and courses are nested in occasions. In such data, students who are nested in courses are nested in occasions. That is, students generally do not repeat the same courses over the semesters. In both data situations, items are crossed with the rest of the data conditions because the same set of items is usually used by students in evaluating teachers and courses over semesters.

Finally, a teacher teaching a course or a course taught by a teacher may be viewed as an undifferentiated instructional event, which represents a third data situation. Thus, three measurement conceptualizations can be formulated to account for the previously described faculty evaluation data that commonly exist. These measurement conceptualizations are discussed next.

Teachers As Objects of Measurement

One conceptualization would have teachers as the objects of measurement. This conceptualization fits the traditional notion of faculty evaluation as is implied by its name. The purpose of such an evaluation is related to promotion and tenure decisions. One is concerned with the dependability[1] of the evaluation over future courses and students that a faculty member might teach and over future evaluations using, possibly, different items. With teachers as the objects of measurement, the rest of the data categorizations represent conditions under which the faculty evaluation has been obtained. In G-theory terminology, a set of similar conditions is called a *facet*. Thus, there are four facets—those of students, courses, items, and occasions —that constitute the universe of admissible observations.

The relations among the evaluation conditions and the objects of measurement are restricted by the existing data structure. For each teacher from a population of teachers, a usual faculty evaluation represents sampling one condition from each of the four facets in the universe of admissible observations. An observed evaluation score of a teacher is a combination of the four sampled conditions. This conceptualization of the faculty evaluation can be summarized in a G-study design of the following form: $(s{:}c){:}(t \times o) \times i$, where an observation, $X_{si:c:to}$, can be decomposed into 11 score effects and the grand mean. These score effects are presented in Table 1.

In Table 1, the grand mean, μ, is the expected value over all teachers in the population of teachers and all the measurement conditions in the universe of admissible observations. The number of teachers in the population and the numbers of conditions in the universe of admissible observations may be considered infinite. The score effects are defined in terms of mean scores. For example, the item score effect is $(\mu_i - \mu)$

[1] The terms *reliability* and *reliability coefficient* are traditionally associated with classical test theory. The analogous terms in generalizability theory are dependability and generalizability coefficient. In this article, however, we use these two sets of terms interchangeably to refer to the same underlying concept of measurement consistency.

TABLE 1
Score Effects and Variance Estimates of the Generalizability Study Designs

	$(s{:}c){:}(t \times o) \times i$ design
Score effects	
$X_{si:c:to}$	$= \mu$
	$+ (\mu_t - \mu)$
	$+ (\mu_i - \mu)$
	$+ (\mu_o - \mu)$
	$+ (\mu_{ti} - \mu_t - \mu_i + \mu)$
	$+ (\mu_{to} - \mu_t - \mu_o + \mu)$
	$+ (\mu_{io} - \mu_i - \mu_o + \mu)$
	$+ (\mu_{tio} - \mu_{ti} - \mu_{to} - \mu_{io} + \mu_t + \mu_i + \mu_o - \mu)$
	$+ (\mu_{c:to} - \mu_{to})$
	$+ (\mu_{ci:to} - \mu_{c:to} - \mu_{tio} + \mu_{to})$
	$+ (\mu_{s:c:to} - \mu_{c:to})$
	$+ (X_{si:c:to} - \mu_{ci:to} - \mu_{s:c:to} + \mu_{c:to})$
Variance components	
$\sigma^2(X_{si:c:to})$	$= \sigma^2(t) + \sigma^2(i) + \sigma^2(o) + \sigma^2(ti) + \sigma^2(to) + \sigma^2(io) + \sigma^2(tio) + \sigma^2(c{:}to) +$ $\sigma^2(ci{:}to) + \sigma^2(s{:}c{:}to) + \sigma^2(si{:}c{:}to)$
Variance estimates	
$\hat{\sigma}^2(t)$	$= [\mathrm{MS}(t) - \mathrm{MS}(ti) - \mathrm{MS}(to) + \mathrm{MS}(tio) + \mathrm{MS}(c{:}to) - \mathrm{MS}(ci{:}to)$ $-\ \mathrm{MS}(s{:}c{:}to) + \mathrm{MS}(si{:}c{:}to]/n_i n_o n_c n_s$
$\hat{\sigma}^2(i)$	$= [\mathrm{MS}(i) - \mathrm{MS}(ti) - \mathrm{MS}(io) + \mathrm{MS}(tio) + \mathrm{MS}(ci{:}to) - \mathrm{MS}(si{:}c{:}to)]/n_t n_o n_c n_s$
$\hat{\sigma}^2(o)$	$= [\mathrm{MS}(o) - \mathrm{MS}(to) - \mathrm{MS}(io) + \mathrm{MS}(tio) + \mathrm{MS}(c{:}to) - \mathrm{MS}(ci{:}to) -$ $\mathrm{MS}(s{:}c{:}to) + \mathrm{MS}(si{:}c{:}to)]/n_i n_t n_c n_s$
$\hat{\sigma}^2(ti)$	$= [\mathrm{MS}(ti) - \mathrm{MS}(tio) + \mathrm{MS}(ci{:}to) - \mathrm{MS}(si{:}c{:}to)]/\mathrm{n}_i n_t n_c n_s$
$\hat{\sigma}^2(to)$	$= [\mathrm{MS}(to) - \mathrm{MS}(tio) - \mathrm{MS}(c{:}to) - \mathrm{MS}(ci{:}to) + \mathrm{MS}(s{:}c{:}to) -$ $\mathrm{MS}(si{:}c{:}to)]/n_i n_c n_s$
$\hat{\sigma}^2(io)$	$= [\mathrm{MS}(io) - \mathrm{MS}(tio) + \mathrm{MS}(ci{:}to) - \mathrm{MS}(si{:}c{:}to)]/n_t n_c n_s$
$\hat{\sigma}^2(tio)$	$= [\mathrm{MS}(tio) - \mathrm{MS}(ci{:}to) + \mathrm{MS}(si{:}c{:}to)/n_c n_s$
$\hat{\sigma}^2(c : to)$	$= [\mathrm{MS}(c{:}to) - \mathrm{MS}(ci{:}to) - \mathrm{MS}(s{:}c{:}to) + \mathrm{MS}(si{:}c{:}to)]/n_i n_s$
$\hat{\sigma}^2(ci : to)$	$= [\mathrm{MS}(ci{:}to) - \mathrm{MS}(si{:}c{:}to)]/n_s$
$\hat{\sigma}^2(s : c : to)$	$= [\mathrm{MS}(s{:}c{:}to) - \mathrm{MS}(si{:}c{:}to)]/n_i$
$\hat{\sigma}^2(sio : c : to)$	$= \mathrm{MS}(si{:}c{:}to)$
	$(s{:}t){:}(c \times o) \times i$ design
	$s{:}(e \times o) \times i$ design
Score effects	
$X_{si:eo}$	$= \mu$
	$+ (\mu_e - \mu)$
	$+ (\mu_i - \mu)$
	$+ (\mu_o - \mu)$
	$+ (\mu_{ei} - \mu_e - \mu_i + \mu)$
	$+ (\mu_{eo} - \mu_e - \mu_o + \mu)$
	$+ (\mu_{io} - \mu_i - \mu_o + \mu)$
	$+ (\mu_{eio} - \mu_{ei} - \mu_{eo} - \mu_{io} + \mu_e + \mu_i + \mu_\mathrm{o} - \mu)$
	$+ (\mu_{s:eo} - \mu_{eo})$
	$+ (X_{si:eo} - \mu_{s:eo} - \mu_{eio} + \mu_{eo})$

(continued)

TABLE 1 (Continued)

Variance components	
$\sigma^2(Xsi{:}eo)$	$= \sigma^2(e) + \sigma^2(i) + \sigma^2(o) + \sigma^2(ei) + \sigma^2(eo) + \sigma^2(io) + \sigma^2(eio) + \sigma^2(s{:}eo) + \sigma^2(si{:}eo)$
Variance estimates	
$\hat{\sigma}^2(e)$	$= [\mathrm{MS}(e) - \mathrm{MS}(ei) - \mathrm{MS}(eo) + \mathrm{MS}(s{:}eo) + \mathrm{MS}(eio) - \mathrm{MS}(si{:}eo)]/n_i n_o n_s$
$\hat{\sigma}^2(i)$	$= [\mathrm{MS}(i) - \mathrm{MS}(ei) - \mathrm{MS}(io) + \mathrm{MS}(eio) - \mathrm{MS}(si{:}eo)]/n_e n_o n_s$
$\hat{\sigma}^2(o)$	$= [\mathrm{MS}(o) - \mathrm{MS}(eo) - \mathrm{MS}(io) - \mathrm{MS}(eio) + \mathrm{MS}(s{:}eo) - \mathrm{MS}(si{:}eo)]/n_e n_i n_s$
$\hat{\sigma}^2(ei)$	$= [\mathrm{MS}(ei) - \mathrm{MS}(eio) + \mathrm{MS}(si{:}eo)]/n_o n_s$
$\hat{\sigma}^2(eo)$	$= [\mathrm{MS}(eo) - \mathrm{MS}(eio) + \mathrm{MS}(s{:}eo) - \mathrm{MS}(si{:}eo)]/n_i n_s$
$\hat{\sigma}^2(io)$	$= [\mathrm{MS}(io) - \mathrm{MS}(eio) + \mathrm{MS}(si{:}eo)]/n_e n_s$
$\hat{\sigma}^2(eio)$	$= [\mathrm{MS}(eio) - \mathrm{MS}(si{:}eo)]/n_s$
$\hat{\sigma}^2(s:eo)$	$= [\mathrm{MS}(s{:}eo) - \mathrm{MS}(si{:}eo)]/n_i$
$\hat{\sigma}^2(si:eo)$	$= \mathrm{MS}(si{:}eo)$

Note. Score effects and variance estimates can be obtained from $(s{:}c){:}(t \times o) \times i$ design by exchanging c with t. s = students; c = courses; t = teachers; o = occasions; e = event.

where μ_i is the mean or expected value of an item over the rest of the conditions in the universe of admissible observations and over the population of teachers. Specifically, $(\mu_i - \mu)$ is the deviation of the mean (over students, courses, occasions, and teachers) associated with an item from the mean over all items, which is the grand mean. Similarly, $(\mu_t - \mu)$ represents the teacher effect, which is the deviation of the mean (over all remaining measurement conditions) associated with a teacher from the mean over all teachers or the grand mean. Another example is the teacher–item interaction effect, $(\mu_{ti} - \mu_t - \mu_i + \mu)$. It is the deviation of the mean (over students within courses and occasions) associated with a teacher and an item from the grand mean after removing the teacher effect and item effect. The last score effect in Table 1 is a residual term that represents a multiway interaction confounded by other sources of variation unaccounted for by the present model.

Squaring and taking the expected value of each score effect yields a variance component. For example, $\sigma^2(i) = E_i(\mu_i - \mu)^2$. Thus, except for the grand mean, which is a constant, 11 variance components are associated with the score effects. One analytical assumption used in G theory is that these score effects are independent of each other. Thus, the 11 variance components are additive. The sum of these components adds up to the total observed score variance.

These variance components, listed in Table 1, are associated with a single teacher in the population and single conditions in the universe of admissible observations. Estimates of these variance components can be obtained from mean squares estimates in an analysis of variance (ANOVA). Adopting a balanced design[2] in which the same number of students and same number of courses are sampled for each teacher, these ANOVA estimates are presented in Table 1.

[2] A nested design is considered balanced if an equal number of conditions of a nested facet is nested in every condition of the nesting facet. For the ease of variance estimation, all the generalizability and

Courses As Objects of Measurement

A different conceptualization of the data would treat courses as the objects of measurement. The purpose of such an evaluation would be to assess the strength and weakness of the course offerings and curriculum. Information from such an evaluation is useful for allocating a budget (Gordon, Jordan, & Albin, 1994) and restructuring courses. One would want to assess the quality and popularity of a course generalizing over the universe of teachers and students who might teach or take it. In this conceptualization, the data categorizations other than courses are treated as conditions imposed to make observations of the objects of measurement. For example, for a course to be evaluated, it has to be taught by a teacher to some students at a particular time point. Thus, the evaluation of the course is subject to the influence of the effectiveness of the teacher as well as the reactions of the students at the time of instruction.

To separate the unique effect of courses independent from that of teachers, only courses that have been taught by more than one teacher can be evaluated. In this situation, teachers are nested within courses. The rest of the data conditions maintain the same relations among themselves as in the first evaluation conceptualization. The G-study design for the current conceptualization is $(s{:}t){:}(c \times o) \times i$. An observed score, $X_{si:t:co}$, is decomposed into 11 score effects as in the previous $(s{:}c){:}(t \times o) \times i$ design. The difference is that, in the current design, the main effect of courses can be estimated, whereas in the $(s{:}c){:}(t \times o) \times i$ design, the course effect is confounded by the interaction between courses and teachers. Because the statistical form of this design is identical to that of the previously discussed $(s{:}c){:}(t \times o) \times i$ design, detailed information on score effects and variance estimates is not provided.

Instructional Events As Objects of Measurement

A third conceptualization would have the combination of a teacher and a course as the object of measurement. This combination has been referred to as an instructional event (Kane et al., 1976). When instructional events are used as the objects of measurement, the course and teacher effects that are undifferentiated are not of interest. Of interest is the effectiveness of an instructional event independent of who teaches what course. Feedback from such an evaluation can be used to determine the instructional quality of a degree program or department. Such feedback is useful for accreditation and institutional accountability (Pratt, 1997; Trow, 1996). Using an instructional event (e) to represent the teacher–course combination, the G-study design for this conceptualization is $s{:}(e \times o) \times i$. The universe of admissible observations con-

decision study designs discussed in this article were balanced, as is the case in almost all published generalizability studies.

sists of three sets of conditions, those of students, items, and occasions. An observed score, $X_{si:eo}$, represents a sampled combination of one student, one item, and one occasion corresponding to an instructional event. $X_{si:eo}$ can be decomposed into nine independent score effects. The associated variance components add up to the total observed score variance. Finally, these variance components can be estimated through a balanced ANOVA. The score effects, variance components, and ANOVA estimates are presented in Table 1.

G-Study Variance Estimates

With the previously described formulations of measurement conceptualizations, the next step in a G study is to take random samples from the universe of admissible observations. The samples are used to estimate the variance components associated with the G-study designs. With respect to faculty evaluation as well as other multifaceted studies where a large data set preexists, random samples from the existing data set can be taken to make the variance estimates.

In this study, samples were taken from the 1995–1996 faculty evaluation at a university in the western United States. For the first conceptualization, $(s{:}c){:}(t \times o) \times i$, where teachers were the objects of measurements, a random sample of 30 teachers ($n_t = 30$) was drawn. For each teacher, two different courses ($n_c = 2$) were sampled for each of two consecutive semesters ($n_o = 2$). The two courses were different across the two semesters. To achieve a balanced design, a random sample of 10 students ($n_s = 10$) from each course was used. Courses that had fewer than 10 students were not included in the sampling frame. A standard evaluation form containing 12 items had been used to obtain all the evaluation data. All 12 items were included in the G study ($n_i = 12$). The items had a 5-point scale describing instructional performance ranging from 1 (*very poor*) to 5 (*very good*).

For the second design, $(s{:}t){:}(c \times o) \times i$, where courses were the objects of measurement, 30 courses ($n_c = 30$) were sampled. Each course was repeated over two semesters ($n_o = 2$). In each semester, two different teachers taught the course ($n_t = 2$). A random sample of 10 students ($n_s = 10$) was used for each course. The same 12 items ($n_i = 12$) were used.

For the third conceptualization, $s{:}(e \times o) \times i$, 30 instructional events were sampled ($n_e = 30$). An instructional event was a combination of a teacher teaching a course. Each instructional event took place for two consecutive semesters ($n_o = 2$). The sampling of the rest of the facets was the same as before. The three samples associated with the three G-study designs were independent.

Variance estimates for all three designs are reported in Table 2. Negative variance estimates were set to zero[3]. In all three designs, the objects of measurement

[3] As a result of sampling variability, variance estimates may become negative even though, by definition, variance components are nonnegative (Brennan, 1983). In the current literature on generalizability theory, negative estimates are set to zero by one of two approaches (Brennan, 1983;

TABLE 2
Random Effects Generalizability Study Estimates of Variance Components

(s:c):(t × o) × i		*(s:t):(c × o) × i*		*s:(e × o) × i*	
Source	$\hat{\sigma}^2$	*Source*	$\hat{\sigma}^2$	*Source*	$\hat{\sigma}^2$
t	.0719	*c*	.1464	*e*	.1122
c:*to*	.0519	*t*:*co*	.1196	*s*:*e*	.4702
i	.0079	*i*	.0089	*i*	.0108
o	.0014	*o*	.0013	*o*	.0023
s:*c*:*to*	.3840	*s*:*t*:*co*	.4442	*ei*	.0155
ti	.0215	*ci*	.0144	*eo*	.0530
to	–.0088	*co*	–.0350	*io*	.0003
io	.0004	*io*	.0001	*eio*	.0147
tio	–.0013	*cio*	.0080	*si*:*eo*	.2883
ci:*to*	.0104	*ti*:*co*	.0254		
si:*c*:*to*	.2736	*si*:*t*:*co*	.2861		

Note. Negative values were set to zero. *s* = students; *c* = courses; *t* = teachers; *o* = occasions; *i* = items; *e* = events.

had a satisfactory amount of variance. This means that the faculty evaluation as conceptualized in this study can differentiate among teachers, courses, or instructional events. It also appears that the components involving facets that were crossed with the objects of measurement were all small. Specifically, in all three designs, the main effects of items and occasions and their interactions with the objects of measurement were small, except for the $\hat{\sigma}^2(eo)$ component in the third design, which was slightly large. The finding regarding the item facet is also consistent with those of three previous studies by Kane et al. (1976), Gillmore et al. (1978), and Smith (1979), respectively, which did not include the occasion facet in the designs.

In all three designs, however, the nested facets had substantial variance components. In the first design, for example, as the largwest variance component,$\hat{\sigma}^2(s:c:to)$was .384, indicating that students' ratings fluctuated greatly over courses and occasions. This nested component, however, contained several confounded effects that, undser the current design, could not be differentiated. Specifically, $\hat{\sigma}^2(s:c:to)$ was confounded by the main effect of students, a two-way interaction between students and courses, and a four-way

Cronbach, Gleser, Nanda, & Rajaratnam, 1972). One approach, which was adopted in this study, is to simply set the negative variance estimate concerned to zero without letting the zero value affect the estimation of other variance components. This approach works when variance estimates are directly derived from the mean squares and sample sizes, as was done in this study. In the other approach, variance estimation is carried out in a hierarchical order where one variance component is derived from estimates of preceding variance components. In this approach, setting one negative estimate to zero is likely to bias other variance estimates.

interaction involving all four components. Among these components, only the last one had a direct bearing on the normative evaluation or rankings of teachers. However, with the existing faculty evaluation data, which were reconceptualized in the form of the (s:c):($t \times o$) $\times$ i design, these components could not be distinctly estimated. To estimate these components, data have to be collected according to designs in which the objects of measurement are fully crossed with all the measurement conditions. Such is the limitation for analyzing existing data sets. On the other hand, most of the previously mentioned data constraints represent inherent characteristics of a faculty evaluation. For example, one cannot expect to have the same students repeat the same courses by the same teachers over time.

In the second design, the objects of measurement, namely, courses, had a larger variance component than in the first design. In Gillmore et al.'s (1978) study, this variance component was close to zero, giving rise to the impression that judgments about courses were undifferentiating (Smith, 1979). However, in replicating Gillmore et al.'s study using a different set of items, Smith obtained a satisfactory variance estimate for courses. Our finding further supports the use of courses as the objects of measurement in faculty evaluation. In this design, the four-way nested term $(\hat{\sigma}^2(s:t:co) = .4442)$ represented, again, the largest variance component. Apart from the imbedded confounding effects, the magnitude of this component implied rating fluctuations on the part of the students. To a lesser degree, the teacher condition $(\hat{\sigma}^2(t:co) = .1196)$ seemed also to introduce inconsistencies in making the evaluation of courses. That is, the evaluation of a course changed depending on who taught it. This interpretation, however, is confounded by the main effect of teachers, which does not bear on the normative evaluation or rankings of courses.

In the third design where instructional events were the objects of measurement, the occasion condition created a larger interaction with the objects of measurement than in the two previously discussed designs. Excluding the nested components that, as in the other two designs, were among the larger components, the interaction between occasion and the objects of measurement $(\hat{\sigma}^2(eo) = .053)$ was the second largest component next to the objects of measurement. In the other two designs, the same interaction component was much smaller. This finding indicates that, when a judgment is made about instructional events independent of who teaches which course, taking evaluations across multiple occasions becomes more important than when teachers or courses are the objects for judgment. In this design, students also represented a slightly larger variance component $(\hat{\sigma}^2(s:eo) = .4702)$ than in the other two designs.

These variance estimates provide useful information for designing efficient faculty evaluation procedures. For example, how many courses should be examined so that a decision regarding faculty promotion achieves a desired level of reliability? This and other D-study considerations are discussed next.

D-STUDY CONSIDERATIONS

A G study is concerned with the conceptualization of a measurement and the estimation of variance components associated with single conditions in the universe of admissible observations. A D study is concerned with efficient applications of the G-study measurement conceptualization and an evaluation of the dependability of these measurement applications. The dependability of a measurement procedure is evaluated in relation to the universe of generalization. A universe of generalization contains all or subsets of the conditions in the universe of admissible observations to which replications of the measurement procedure are to be generalized. In other words, the universe of generalization sets the scope within which dependability of the measurement procedure is evaluated. For example, an institution that uses only one set of items (i.e., a standard evaluation form) in conducting faculty evaluation would not be concerned with the consistency of the evaluation in terms of other items or evaluation forms. However, it may be concerned with the dependability of the evaluation when it is used to assess future courses that the faculty members might be teaching. The universe of generalization in this example contains the course facet but not the item facet of the universe of admissible observations.

When one uses faculty evaluations to make decisions (e.g., who will receive a promotion, which course should cease to be offered, or what instructional program is most effective), total scores or mean scores based on multiple measurement conditions will provide more dependable results than single scores. Thus, a universe of generalization is normally associated with mean scores over multiple conditions sampled from the universe of admissible observations. Based on G-study variance estimates, which are associated with single conditions, one decides, in a D study, on an efficient number of conditions to be sampled to achieve a prescribed level of dependability. The purpose of this D study is to determine how many measurement conditions from the existing faculty evaluation data should be used in making different evaluation decisions. In the following, within each of the three conceptualizations of faculty evaluations, three D-study considerations are provided that make efficient use of the existing data for various decision-making purposes. These D-study considerations are based on the respective G-study data reported earlier, rather than on new data. Thus, they are called *D-study considerations* instead of D studies (Brennan, 1983).

Teacher-Related Considerations

The three D-study considerations discussed in this section are associated with the first G-study design, $(s{:}c){:}(t \times o) \times i$, where teachers were the objects of measurement. In one decision consideration, it is reasonable to define the universe of generalization as being identical to the universe of admissible observations. That is, the decision maker is interested in generalizing to all the facets when considering the dependability of a faculty evaluation. It is also assumed that the decision maker is interested in the mean scores of teachers over multiple conditions from each facet. The D-study design will be (S:C):(t × O) × I, where a capital letter indicates the mean over multiple conditions of a facet. The observed mean score of a teacher is $\bar{X}_t$ (or, $X_{SI:C:tO}$). This mean score is the result of one instance of the evaluation procedure that samples a combination of n'_s students, n'_i items, n'_ccourses, and n'_o occasions from the universe of admissible observations. Another instance of the same evaluation procedure will result in sampling another combination of n'_s , n'_i , n'_c ,and n'_o conditions from the same universe of admissible observations. The set of all combinations of n'_s , n'_i , n'_c ,and n'_o conditions constitutes the universe of generalization. In making dependability forecasts, a D-study sample size, n'_a , does not have to be the same as, and is usually different, from a G-study sample size, n_a.

In this D-study consideration, a teacher’s universe score is the expected value over the defined universe of generalization. It is defined as $\mu_t \equiv E_I E_O E_C E_S \bar{X}_t$ (or $\mu_t \equiv E_I E_O E_C E_S X_{SI:C:tO}$). Universe score variance is $\sigma^2(t) = E_t(\mu_t - E_t\mu_t)^2$. The expected observed score variance (the variance of teachers’ observed mean scores) is:

$$\begin{aligned} E\sigma^2(\bar{X}_t) &\equiv E_t E_I E_O E_C E_S (X_{SI:C:tO} - E_t X_{SI:C:tO})^2 \\ &= \sigma^2(t) + \sigma^2(It) + \sigma^2(to) + \sigma^2(tIO) + \sigma^2(C:tO) + \\ &\quad \sigma^2(CI:tO) + \sigma^2(S:C:tO) + \sigma^2(SI:C:tO). \end{aligned}$$

Variance components due to items ($\sigma^2(I)$), occasions ($\sigma^2(O)$), and the interaction between the two ($\sigma^2(IO)$) are not included in the expected observed score variance of teachers because these effects are constant to all teachers. For example, if difficult items are sampled, they affect the evaluation of all teachers. These components are distinguished from the interaction components involving teachers, $\sigma^2(tI)$ and $\sigma^2(tIO)$, which are included in the expected observed score variance because they affect the relative standings of teachers.

The G coefficient is defined as the ratio of universe score variance to the expected observed score variance, $E\rho^2(CSIO) = \sigma^2(t) / E\sigma^2(\bar{X}_t)$. Capital letters in the parentheses indicate the facets to which the estimated G coefficient is intended to generalize. Specifically, $E\rho^2(CSIO)$ indicates the degree to which, under the current D-study consideration, an observed score, $\bar{X}_t$, can be replicated from similar mean scores of other randomly sampled n'_c courses, n'_s students, n'_i items, and n'_o occasions. The sample size is part of the D-study consideration. Figure 1, which will be discussed at the end of the D-study section, plots the changes in G-coefficient estimates as a function of different sample size considerations.

$E\rho^2(CSIO)$ is approximately the expected value of the squared correlation of two evaluations based on random samples of n'_s students within n'_c courses and n'_o occasions using n'_i items. One can also view $E\rho^2(CSIO)$ as analogous to a reliability coefficient in classical test theory. The expected observed score variance consists of two parts, the universe score variance, $\sigma^2(t)$, and variance due to different conditions on which the measurement is taken. In G theory, these latter variance components constitute $\sigma^2(\delta)$, which is called error variance for a norm-referenced interpretation (Brennan, 1983). In this light, the G coefficient can be interpreted in a similar fashion to a reliability coefficient in classical test theory except that the error variance here contains multiple sources representing different conditions on which the measurement is taken. Taking the square root of $\sigma^2(\delta)$ yields what is conceptually equivalent to the standard error of measurement in classical test theory.

Whereas $E\rho^2(CSIO)$ generalizes to all the facets in the universe of admissible observations, another D-study consideration could have the item facet fixed. Fixing a facet means that the same sampled conditions from the facet will be retained in future replications of the measurement procedure. Fixing the item facet is reasonable because most institutions use one set of items or a standard form of evaluation throughout the years. With items fixed, the universe of generalization is smaller in that it does not have the item facet to which to generalize replications of the evaluation procedure. The intention is not to generalize to other items but to use, in future evaluations, the same items that are included in the D study.

When items are fixed, the universe score is the expected value of the observed mean score over one fixed set of n'_i items and all possible samples of n'_s students, n'_c courses, and n'_o occasions. The universe score variance is $\sigma^2(t) + \sigma^2(tI)$. Because the expected value is not taken over all possible items in the universe of admissible observations, the systematic effect due to the particular set of items will become part of the measurement in future replications. Thus, $\sigma^2(tI)$ becomes part of the systematic universe score variance. The expected observed score variance remains unchanged. The G coefficient for generalizing over students, courses, occasions, but not items is $E\rho^2(CSO) = [\sigma^2(t) + \sigma^2(tI)] / E\sigma^2(\bar{X}_t)$. $E\rho^2(CSO)$ is approximately the expected value of the squared correlation of two evaluations based on the same n'_i items and random samples of n'_s students, n'_c courses, and n'_o occasions. This G coefficient would be larger than the previously discussed G coefficient, $E\rho^2(CSIO)$, because the universe of generalization is smaller. The measurement procedure has higher generalizability with restricted generalization. Figure 1 illustrates the changes in G-coefficient estimates as a function of sample size changes.

A third D-study consideration could have both items and courses fixed. Fixing courses is reasonable because many faculty members repeatedly teach a few courses that may be included in a D study. Fixing courses means future evaluation of a faculty will be based on the same courses that are used in the D study. Thus,

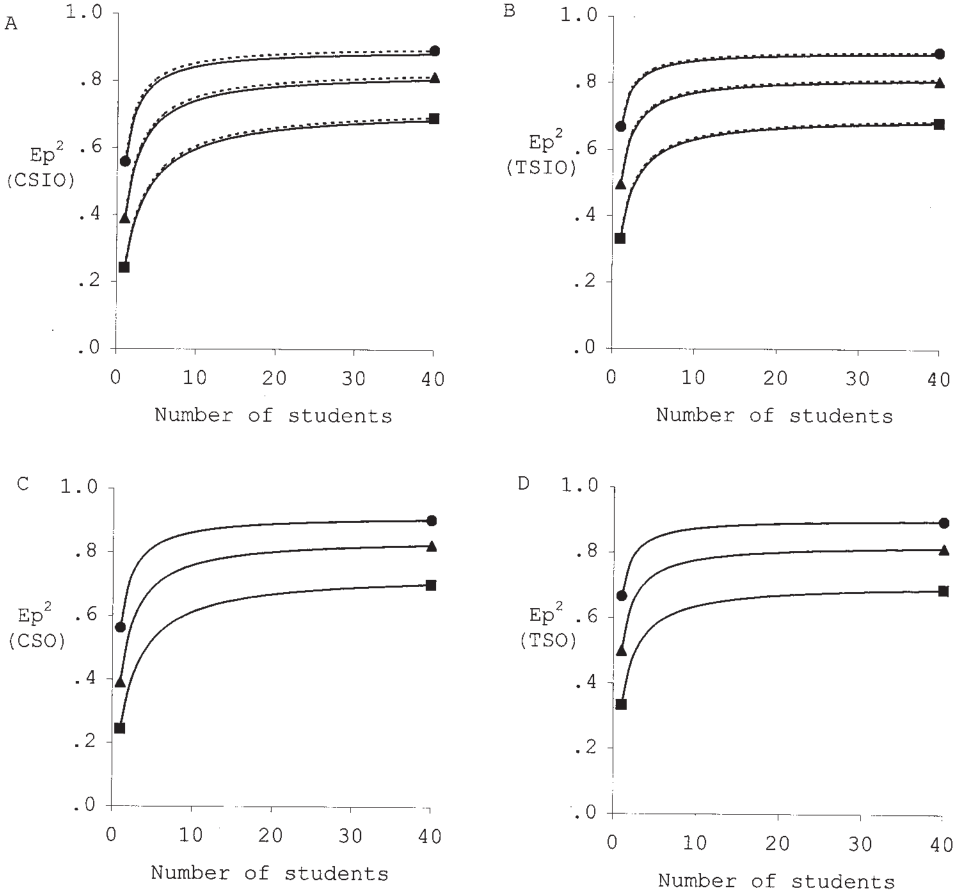

FIGURE 1 (Continued)

variance resulting from the teacher–course interaction (i.e., variability due to a teacher performing differently on assigned courses) becomes part of the universe score variance. However, with the current design, where courses were nested in both teachers and occasions, the teacher–course interaction and the three-way teacher–course–occasion interaction are inseparable. They are imbedded in the $\sigma^2(C{:}tO)$ component. By fixing courses, the whole component of $\sigma^2(C{:}tO)$ has to become part of the universe score variance. In this approach, the effect of occasion on course is ignored. In other words, the part of the teacher–course fluctuation that is also a result of teaching the courses at different times is ignored. Consequently, the estimate of the G coefficient is biased upward. Because items are also fixed, the same explanation applies to the treatment of $\sigma^2(CI{:}tO)$ as part of the universe score variance. (The approach, however, is different than fixing the occasion facet. If the

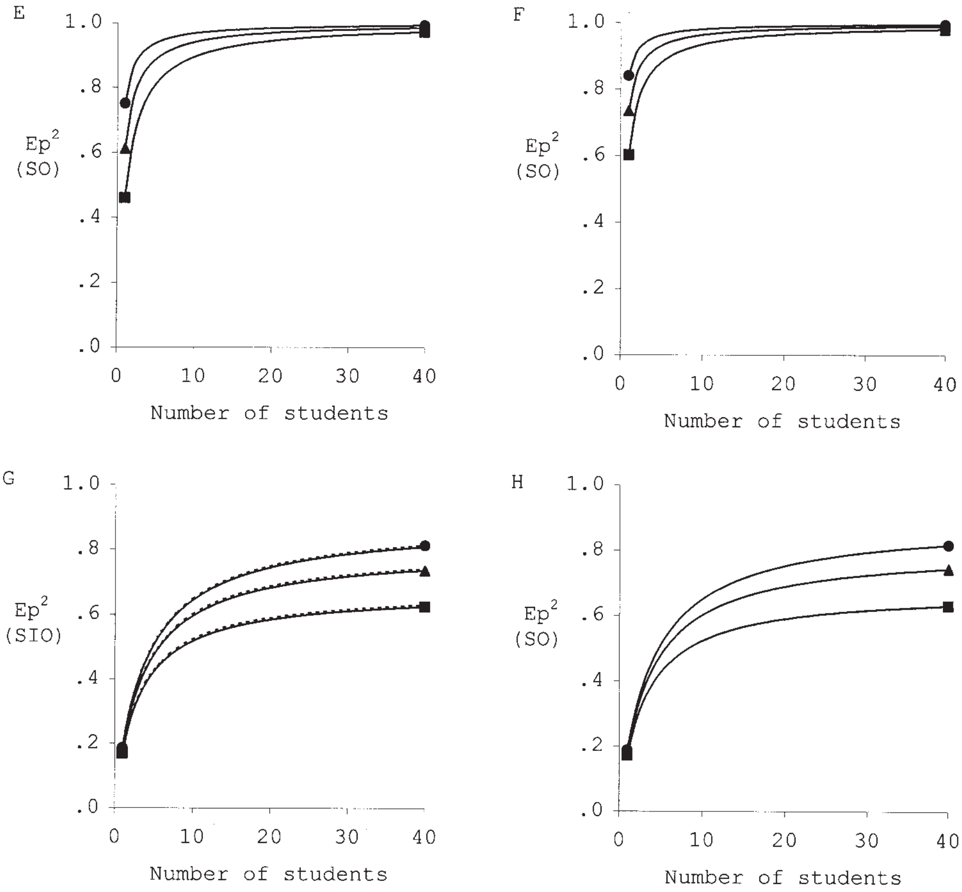

FIGURE 1 Generalizability- (G-) coefficient estimates under different sample size considerations. In Plot A, teacher = object of measurement; occasion = 2; item = 12 for solid line, 24 for dotted line; course = 1, 2, 4, for square, triangle, and circle. In Plot B, course = object of measurement; occasion = 2; item = 12 for solid line, 24 for dotted line; teacher = 1, 2, 4 for square, triangle, and circle. In Plot C, teacher = object of measurement; occasion = 2; item = fixed; course = 1, 2, 4, for square, triangle, and circle. In Plot D, course = object of measurement; item = fixed; occasion = 2; teacher = 1, 2, 4 for square, triangle, and circle. In Plot E, teacher = object of measurement; item = fixed; course = fixed; occasion = 1, 2, 4 for square, triangle, and circle. In Plot F, course = object of measurement; item = fixed; teacher = fixed; occasion = 1, 2, 4 for square, triangle, and circle. In Plot G, instructional event = object of measurement; item = 12 for solid line, 24 for dotted line; occasion = 1, 2, 4 for square, triangle, and circle. In Plot H, instructional event = object of measurement; item = fixed; occasion = 1, 2, 4 for square, triangle, and circle.

occasion facet also were fixed, two more components, $\sigma^2(tIO)$ and $\sigma^2(CI{:}tO)$, would be included in the universe score variance.) With items and courses both fixed, the universe score variance is $\sigma^2(t) + \sigma^2(tI) + \sigma^2(C{:}tO) + \sigma^2(CI{:}tO)$. The expected observed score variance is unchanged. The G coefficient for generalizing over students and occasions but not items or courses is $E\rho^2(SO) = [\sigma^2(t) + \sigma^2(tI) + \sigma^2(C:t) + \sigma^2(CI:t)]/E\sigma^2(\bar{X}_t)$. G-coefficient estimates associated with different sample size considerations are reported in Figure 1.

Course and Program Related Considerations

Similar D-study considerations are made for the other two G-study designs. One of the G-study designs, $(s{:}t){:}(c \times o) \times i$, treats courses as the objects of measurement. The expected observed score variance of courses is:

$$\begin{aligned} E\sigma^2(\bar{X}_c) &= E_c E_I E_O E_T E_S (X_{SI:T:cO} - E_c X_{SI:T:cO})^2 \\ &= \sigma^2(c) + \sigma^2(cI) + \sigma^2(cO) + \sigma^2(cIO) + \sigma^2(T:cO) + \\ &\quad \sigma^2(TI:cO) + \sigma^2(S:T:cO) + \sigma^2(SI:T:cO). \end{aligned}$$

The G coefficient for generalizing over all measurement conditions is:

$$E\rho^2(TSIO) = \sigma^2(c)/E\sigma^2(\bar{X}_c)\ .$$

The G coefficient for generalizing over teachers, students, and occasions but not items is:

$$E\rho^2(TSO) = [\sigma^2(c) + \sigma^2(cI)]/E\sigma^2(\bar{X}_c)\ .$$

The G coefficient for generalizing over students and occasions but not over items or teachers is:

$$E\rho^2(SO) = [\sigma^2(c) + \sigma^2(cI) + \sigma^2(T:cO) + \sigma^2(TI:cO)]/E\sigma^2(\bar{X}_c)\ .$$

In the other design, $s{:}(e \times o) \times i$, instructional events are the objects of measurement. The expected observed score variance is:

$$\begin{aligned} E\sigma^2(\bar{X}_e) &\equiv E_e E_I E_O E_S (X_{SI:eO} - E_e X_{SI:eO})^2 \\ &= \sigma^2(e) + \sigma^2(eI) + \sigma^2(eO) + \sigma^2(eIO) + \\ &\quad \sigma^2(S:eO) + \sigma^2(SI:eO). \end{aligned}$$

The G coefficient generalizing over all conditions is:

$$E\rho^2(SIO) = \sigma^2(e) / E\sigma^2(\bar{X}_e).$$

The G coefficient generalizing over students and occasions but not items is:

$$E\rho^2(SO) = [\sigma^2(e) + \sigma^2(eI)] / E\sigma^2(\bar{X}_e).$$

The G coefficient for generalizing over students only but not over items or occasions is:

$$E\rho^2(S) = [\sigma^2(e) + \sigma^2(eI) + \sigma^2(eO) + \sigma^2(eIO)] / E\sigma^2(\bar{X}_e).$$

Estimates of the previously discussed G coefficients can be directly derived from the estimates of the universe score variance and different D-study variance components. The latter are obtained in the same manner as are the G-study variance components described in Table 1. According to the central limit theorem, the variance of mean scores is that of individual scores divided by the sample size. If the D study is based on the same data set as the G study, as is the case with this study, D-study variance estimates can be directly obtained by dividing the corresponding G-study estimates by the number of conditions sampled in the D study. Estimates of different G coefficients associated with different sample size considerations are plotted in Figure 1.

Sample Size Recommendations

One of the common D-study considerations is to determine, for each set of measurement conditions, an efficient sample size that maximizes the dependability of the measurement procedure while minimizing costs. Figure 1 plots the changes in G-coefficient estimates as a function of the different sample size changes. The increments in sample size associated with different sets of measurement conditions in Figure 1 are proportional to their original G-study sample sizes. For example, an increase of two courses (which represented a 100% increase over its original G-study sample size of two) was compared to an increment of 12 item conditions (which is also a 100% increase over its G-study sample size of 12). Thus, one can draw comparisons of the relative impact of sample size increments on the G coefficients across the different sets of measurement conditions.

In all three designs using teachers, courses, and instructional events as the objects of measurement, increasing evaluation items had little impact on G estimates.

As shown in Plots A, B, and G of Figure 1, the dotted line that represents 24 items overlaps the solid line that represents 12 items. Thus, other conditions being equal, using 12 items achieved almost the same level of dependability as using 24 items. Similarly, fixing items did not attain as much of a dependability increment as this approach was intended to achieve. This conclusion is drawn from the small difference between Plots A and B, where items were random, and Plots C and D, where items were fixed. This is especially true for the first two designs. For example, calculated from the G-coefficient estimates that were used to plot Figure 1, estimates of the two G coefficients associated with sampling 4 teachers, 20 students, 12 items, and 4 occasions were $E\hat{\rho}^2(TSIO) = .934$ for random items and $E\hat{\rho}^2(TSO) = .941$ with the 12 items fixed. Gillmore et al. (1978) reported equally similar G coefficients for generalizing versus not generalizing over the item facet. An implication of the finding is that universities that use different evaluation items, either across time or across academic disciplines, may still draw dependable comparisons of the rankings of teachers or courses.

As shown in almost all of the plots in Figure 1, an efficient number of students to be sampled in conducting evaluations seemed to be between 10 and 20. Above 20 students, the increment on measurement dependability became decreasingly small. Similar findings were reported by Gillmore et al. (1978) and Kane et al. (1976). More important, however, the measurement dependability of the evaluation decelerated fast when the number of student conditions went below 10.

When teachers were the objects of measurement, increasing courses seemed to have the highest impact on generalizability estimates. As shown in Figure 1, an efficient number of courses seemed to be between two and four. Although adding courses increased dependability, using more than four courses was clearly not cost effective. This finding is different than that of Gillmore et al. (1978) who recommended sampling 5 to 10 courses. As shown in Figure 1, with a sufficient number of students, the evaluation based on a single course had an acceptable generalizability of .60 or higher. When courses were the objects of evaluation, the sample size of teachers teaching the same courses had slightly more impact on the G-coefficient estimate than the sampling of courses when teachers were the objects of measurement. An efficient sample size also seemed to lie between two and four. When courses (or teachers) and items were both fixed, that is, the same items and courses (or teachers) were used to evaluate a teacher (or a course), collecting such evaluation data once, twice, or four times seemed to make little difference on the dependability of the results. This can be seen from Plots E and F of Figure 1. In other words, for example, if one is only interested in how well a teacher teaches the same course, the results based on one evaluation or the average of two or four such evaluations were almost equally highly dependable, given that the same items and a sufficient number of students were used. However, the occasion condition made a much larger impact when instructional events were the objects of evaluation. This finding is not surprising because the ratings of instructional events were confounded by both the teacher and course effect and, thus, should be less consistent across semesters.

DISCUSSION

In this study, we demonstrate the use of G theory in analyzing an existing set of faculty evaluation data. When faculty evaluation is initially conducted, there is usually not a G-theory conception of what constitutes the universe of admissible observations. When a substantive decision is to be made or has been made, one may want to know the dependability of that decision. As shown in this article, one can reconceptualize the existing evaluation data within the framework of G theory. Random samples can then be taken from existing data to estimate the variance components associated with the conceptualization. These variance estimates enable one to choose an efficient D-study design from which substantive decisions can be made with a known level of dependability.

For example, to achieve the same level of dependability, one may sample fewer courses when making decisions regarding a faculty member teaching the same courses than teaching different courses. Because a university is usually concerned with how well a faculty member teaches the same set of courses she or he normally teaches, the university may not need to conduct teaching evaluations for every course the faculty member teaches in a semester. Instead, a random sample of courses can be selected from a faculty member for evaluation. The sample size can be as low as one course because, as shown in the study, teaching evaluations based on a single course had an acceptable reliability of .60 or higher. Sometimes, a professor is recognized for excellence in teaching from the teaching of a single course. In light of these findings, such a teaching reputation can be perceived as sustainable and generalizable, given that enough items are used by a sufficient number of students in providing the evaluation of the course. When generalizing over all possible courses, an average rating of a faculty member based on two to four courses seems to be desirable. An evaluation based on more than four courses is clearly not cost effective.

The same is true for making decisions regarding a course as the object of judgment. First, it is sufficient to look at the average rating of a course based on two to four different teachers, if any teacher is expected to teach the course. When teachers are fixed, the judgment about the course being taught by the same teachers is, not surprisingly, more dependable. When items were fixed, the average dependability of a decision regarding a course to be taught by any teacher was .827. The average dependability of a judgment of the course to be taught by the same teachers who were sampled in a D study was .964. Thus, when decisions are made about courses, for example, in curriculum evaluation, course restructuring, and allocation of resources (Gordon et al., 1994), one does not need to sample many different teachers teaching a course, especially when the same teachers are expected to continue to teach the course.

The impact of the number of students on teaching evaluation is an interesting one. One inflection point of the dependability curve, as shown in Figure 1, seemed to be around 10 students, below which the deterioration of reliability accelerated quickly. This finding is important because it is common to hear one or two students comment on a faculty member’s teaching. Such comments, which tend to be clearly positive or negative, are sometimes taken seriously by an academic department or even higher levels of the university administration. As shown in Figure 1, the evaluation by one or two students was highly unreliable. Such an evaluation could be replicated as little as 20% of the time, even when an adequate number of items was used. Another implication of the same finding is that measurement dependability could be questionable for the evaluation of an extremely small class with, for example, five to six students.

An acceptable number of students seems to be between 10 and 20. Above the sample size of 20, however, the number of students seems to have little impact on the consistency of an evaluation, when either the teachers, courses, or, to a lesser degree, instructional events are the objects of judgment. For example, the average of the G-coefficient estimates based on a 10-student teacher evaluation ($E\hat{\rho}^2(CSIO) = .777$) was almost identical to that of a 20-student evaluation ($E\hat{\rho}^2(CSIO) = .792$), averaging over other conditions. However, this finding does not imply that class size has no bearing on faculty evaluation. That is, independent of the consistency of the evaluation, which up to a certain point does not seem to be influenced by the number of students in a class, it is not known, for example, whether small classes tend to provide systematically more positive evaluations than large classes. Future studies can explore this question by including class size as an additional measurement condition. The findings discussed here are based on the assumption that all the measurement conditions were randomly sampled from the universe of admissible observations. If, for example, certain students choose not to evaluate a teacher for some reason, the reduced sample size of the student conditions would, of course, have different implications.

Instructional events as the objects for decision making have been rarely used in the literature. However, with the current focus on institutional accountability (Trow, 1996), it becomes increasingly important to evaluate instructional events independent of who teaches what course. Accreditation, for example, calls for an evaluation of course offerings without too much concern over what courses are taught by which teachers. Similar evaluations are relevant for fund-raising, budgeting, and public relations purposes (Gordon et al., 1994). A teacher and a course that form an instructional event, individually, no longer represent measurement conditions. As shown in the D-study results, the potential concern for judgment dependability seems to lie in the number of semesters for which a decision should be based. Such a decision, as shown in Figure 1, should be based on the average rating of at least two semesters of instruction.

Because most universities have extensive faculty evaluation data, one can make improvements over this study by taking multiple samples from an existing data set. An average of multiple estimates of the same G-study variance components can be used in subsequent decision considerations. The mean estimate will have higher precision and stability than a single estimate. In addition, one can obtain an estimate of the standard error of the mean variance estimates. This approach has been demonstrated by Brennan, Gao, and Colton (1995).

The three G-study conceptualizations considered here are representative of the common data structure and use of faculty evaluation. Variance estimates from this G study can be used by decision makers to evaluate the dependability of different D-study decisions without conducting another G study. As long as the measurement conceptualization is the same, variance estimates from one G study can be used for multiple D studies carried out by different people. However, conceptualizations different from what are presented in this study may also be formulated to account for an existing evaluation. For example, in this study, all facets were considered infinite, whereas it is also reasonable to conceptualize the course facet as finite because most universities list a fixed number of courses to be offered. With finite courses, the designs and their interpretations will remain the same as those presented in this study except that the concept of taking expected values over a random facet will be replaced by averaging over the *N* levels of the finite facet. Variance estimation can also remain the same as that presented in Table 1, plus the application of the finite universe correction factor (Cochran, 1977) to the components involving the finite facet as a nesting factor. The correction is the reciprocal of $1 - n / N$, where n is the sampled number of conditions from the finite facet and N is the finite number of conditions of the facet in the universe.

ACKNOWLEDGMENTS

This article was partly supported by Earmarked Grant 4047/98H from the Research Grant Council of the Hong Kong government. A previous version of this article was presented at the 1998 annual convention of the American Educational Research Association, San Diego, CA. We thank James Impara and three anonymous reviewers for their helpful comments on earlier drafts of this manuscript.

REFERENCES

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

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11,* 27–34.

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of work keys listening and writing tests. *Educational and Psychological Measurement, 55,* 157–176.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

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.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement, 15,* 1–13.

Gordon, G., Jordan, C. E., & Albin, M. J. (1994). Accountability and academic improvement on a departmental level: One school's approach. *Journal of Education for Business, 69,* 288–291.

Kane, M. T., & Brennan, R. L. (1977). The generalizability study of class means. *Review of Educational Research, 47,* 267–292.

Kane, M. T., Gillmore, G. M, & Crooks, T. J. (1976). Student evaluations of teaching: The generalizability study of class means. *Journal of Educational Measurement, 13,* 171–183.

Pratt, D. D. (1997). Reconceptualizing the evaluation of teaching in higher education. *Higher Education, 34,* 23–44.

Shavelson, R. J., & Webb, N. M. (1988). Generalizability theory. *American Psychologist, 44,* 922–932.

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

Smith, P. L. (1979). The generalizability of student ratings of courses: Asking the right questions. *Journal of Educational Measurement, 16,* 77–87.

Trow, M. (1996). Trust, markets and accountability in higher education: A comparative perspective. *Higher Education Policy, 9,* 309–324.