Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 51

Using G Theory to Examine Confounded Effects: "The Problem of One"

Robert L. Brennan[1]

January 2017

[1] Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Founding Director, Center for Advanced Studies in Measurement and Assessment (CASMA). Brennan is also a consultant to the College Board.

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: https://education.uiowa.edu/centers/casma

Contents

Abstract

Using the conceptual framework of G theory, this paper addresses the issue of confounded effects that arise when the data collection design for an operational measurement procedure has a single condition for one or more facets. In this case, effects for at least two facets will be confounded, which leads to ambiguities in interpreting results. We call this the "the problem of one." Ambiguities are particularly salient when one of the confounded effects is random and one is fixed. Confounding arises frequently in performance testing (e.g., essay testing), although confounding tends to be neglected or, worse still, unrecognized.

Initial attention focuses on using a single prompt such as an essay, followed by brief discussions of important matters that relate to confounding, in general. Then consideration is given to designs in which single prompts are completely confounded with multiple fixed prompt types. The third major section considers confounding associated with using a single "rater" such as an automated scoring engine.

This paper is focused mainly on conceptual issues and how single conditions of facets almost always lead to bias in error variances and coefficients. Understanding of these issues necessitates careful distinctions between: (a) the universe of admissible observations and the actual measurement procedure; and (b) which facets are fixed and which are random. In a sense, this paper challenges certain aspects of the "conventional wisdom" in psychometrics

Generalizability (G) theory offers an extensive conceptual framework and a powerful set of statistical procedures for addressing numerous measurement issues. The defining treatment of G theory is a monograph by Cronbach, Gleser, Nanda, and Rajaratnam (1972). Brennan (2001) provides the most extensive current treatment of G theory. It is assumed here that readers have some familiarity with G theory.

The paper addresses the issue of confounded effects in G theory that arise when a data collection design for an operational measurement procedure has a single condition for one or more facets. In this case, effects for at least two facets will be confounded, which leads to ambiguities in interpreting results. We call this the "the problem of one." Ambiguities are particularly salient when one of the confounded effects is random and one is fixed. Confounding arises frequently in performance testing (e.g., essay testing), although confounding tends to be neglected or, worse still, unrecognized.

In their original treatment of G theory, Cronbach et al. (1972) acknowledged the issue of confounded effects and even routinely used a relatively complex notational system to denote them. Brennan (2001, see especially pp. 62–63) also deals with confounded effects. Both books, however, treat confounded effects primarily from the point of view of identifying the effects that are confounded in a nested effect.[2] Here, we are considering confounded effects that arise primarily because of single conditions of facets in an actual measurement procedure. This is closely associated with the issue of hidden facets (see, for example, Brennan (2001, pp. 149–153), but the treatment of confounded effects in this paper is much more extensive than that in Cronbach et al. (1972) or Brennan (2001).

This paper is focused mainly on conceptual issues and how single conditions of facets almost always lead to bias in error variances and coefficients. Understanding of these issues necessitates careful distinctions between: (a) the universe of admissible observations (UAO) and the actual measurement procedure; and (b) which facets are fixed and which are random.

Failure to recognize these issues is much more pervasive that is generally acknowledged, which causes considerable ambiguity in the interpretation of many psychometric analyses. It is important to note, however, that G theory does not introduce or cause these ambiguities; rather, G theory provides the only current framework for coherently examining these issues.

Section 1 focuses on confounding associated with using a single prompt such as an essay. Section 1 also includes discussions of important matters that relate to confounding, in general. Section 2 extends Section 1 to designs that involve multiple prompt types. Section 3 focuses primarily on confounding associated with using a single rater (e.g., and automatic scoring engine). The distinction between a single prompt and a single rater is somewhat artificial since both of them are facets, and many of the conceptual and technical issues associated with a single prompt could apply to a single rater, and vice-versa. Since confounding is such a challenging topic, however, it is useful to discuss matters in familiar

[2] For example, if i is crossed with h in the universe (i.e., $i \times h$), but i is nested within h in a design (i.e., $i{:}h$), then we say that $i{:}h$ involves the confounding of i and ih from the crossed universe.

contexts. (Gao, Brennan, & Guo, 2015, consider some types of confounding that occur with single conditions of an occasion facet, as does Brennan, 2001, pp. 127–128.)

1 A Single Prompt

We begin with a seemingly simple example for a single prompt that turns out to be more complex than might be expected. Then, in Section 2 we turn to a more complicated example in which prompt and prompt-types are confounded.

1.1 The $p \times r$ Design with a Single Prompt

Suppose the UAO has two crossed facets, raters (r) and prompts (z), that are crossed with persons. Further, suppose that $p \times r \times z$, with the population of persons being large and the two facets having large numbers of conditions. By contrast, suppose the G study data are collected using multiple persons and multiple raters, but only *one* prompt. The G study data, then, constitute a $p \times r$ design, but the data are completely blind to the existence of multiple prompts in the UAO. In fact, the single prompt is hidden in the data (i.e., confounded with every single data element), which means that any analysis of the data effectively treats prompt as a *single fixed* condition.

A typical G theory analysis of the data for the $p \times r$ design would yield estimates of three variance components that are usually designated $\sigma^2(p)$, $\sigma^2(i)$, and $\sigma^2(pi)$. Then, typically, various D study statistics would be computed, such as relative error variance and a generalizability coefficient:

$$\sigma^2(\delta) = \frac{\sigma^2(pr)}{n'_r} \tag{1}$$

and

$$\boldsymbol{E}\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\delta)}. \tag{2}$$

For the UAO considered here, however, these representations of $\sigma^2(\delta)$ and $\boldsymbol{E}\rho^2$ are quite misleading, because they do not reflect the influence on the G study data of the single level of the z facet. In this case, therefore, a more descriptive version of Equations 1 and 2 is

$$\sigma^2(\delta) = \frac{\sigma^2(pr|z)}{n'_r} \tag{3}$$

and

$$\boldsymbol{E}\rho^2 = \frac{\sigma^2(p|z)}{\sigma^2(p|z) + \sigma^2(\delta|z)}, \tag{4}$$

where "$|z$" designates that there is a single *fixed* level of z in the G study data.

Using procedures discussed in Brennan (2001, pp. 120–124), it follows that

$$\sigma^2(\delta) = \frac{\sigma^2(pr) + \sigma^2(prz)}{n'_r} \tag{5}$$

and

$$\boldsymbol{E}\rho^2 = \frac{\sigma^2(p) + \sigma^2(pz)}{\sigma^2(p) + \sigma^2(pz) + \sigma^2(pr)/n'_r + \sigma^2(prz)/n'_r}, \tag{6}$$

where the variance components in Equations 5 and 6 are for the UAO in which z is a random facet.

Equations 5 and 6 are misleading relative to the UAO, however. In particular, since the investigator has specified that the z facet is *random* in the UAO, and since $n_z = n'_z = 1$, the equations for $\sigma^2(\delta)$ and $\boldsymbol{E}\rho^2$ that reflect the investigator's *intent* are

$$\sigma^2(\delta) = \sigma^2(pz) + \frac{\sigma^2(pr) + \sigma^2(prz)}{n'_r} \tag{7}$$

and

$$\boldsymbol{E}\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(pz) + \sigma^2(pr)/n'_r + \sigma^2(prz)/n'_r}. \tag{8}$$

Clearly, $\sigma^2(\delta)$ in Equation 5 will be an *under*statement of the intended value given by Equation 7. By contrast, $\boldsymbol{E}\rho^2$ in Equation 6 will be an *over*statement of the intended value given by Equation 8. In short, the G study $p \times r$ design based on a single condition of the z facet leads to bias in error variances and coefficients. The fundamental cause of this bias is that the G study $p \times r$ design has a single fixed level of z, which makes it impossible to disentangle $\sigma^2(p)$ and $\sigma^2(pz)$.

1.2 Complexities and Potential Misunderstandings

In the above discussion and throughout this paper, careful attention needs to be given to various issues, complexities, and potential misunderstandings that can arise when confounding is present. Again, these issues arise essentially because there is a mismatch between a UAO and the design of an operational measurement procedure.

1.2.1 Structural vs. Statistical Bias

In traditional statistics and psychometrics literature, almost always "bias" refers to bias in numerical estimates of parameters caused by the use of certain estimators that are known to have the property that the expected value of the estimator does not equal the parameter (e.g., Bayesian estimators are usually biased).[3] In this paper, any reference to bias has nothing to do with statistical bias.

[3] As discussed by Brennan (2001), in G theory the usual ANOVA estimators of variance components are unbiased.

Rather, here bias is to be understood in a structural sense caused by the fact that the design of an operational measurement procedure fails to mirror the intended UAO. That is, the design of the operational measurement procedure is

- narrower that the UAO and/or
- ambiguous with respect to one or more characteristics of the UAO.

When such a mismatch occurs between the UAO and the measurement procedure, we sometimes refer to it as ”bias,” but we avoid using the phrase “biased estimates” because it is so closely tied to notions of statistical bias. For the same reason, we sometimes use terms such as “overstate” or “understate,” rather than “overestimate” or “underestimate.”

There are similarities between results for the confounding discussed in this section and Kane's (1982) consideration of the reliability-validity paradox, as summarized by Brennan (2001, pp. 132–135). For example, in Kane's framework $\boldsymbol{E}\rho^2$ in Equation 6 is for a restricted universe and $\boldsymbol{E}\rho^2$ in Equation 8 is for an unrestricted (or target) universe of generalization. Kane's treatment, however, does not focus primarily on confounding as discussed in this paper. This paper and Kane's are not contradictory, however; indeed, they are more properly viewed as complementary.

1.2.2 Notational and Verbal Complexity

Confounding is a conceptually challenging topic that is very difficult to represent fully and unambiguously in notation and/or words, without using expressions that can be complicated or confusing. In this paper, we usually opt for expressions that are reasonably simple, even though such expressions could be misunderstood if taken out of context.

For example, it was stated above that:

> It is evident that, using the G study data, $\sigma^2(\delta)$ in Equation 5 will be an *under*statement of the intended value given by Equation 7.

Strictly speaking, the G study *data* do not provide the *parameters* in Equation 5. We could say something like “the G study data can be used to estimate $\sigma^2(\delta)$ in Equation 5,” but then the word “estimate” would likely trigger notions of statistical basis, which is not a matter discussed in this paper. The above statement is also a bit misleading because the $\sigma^2(\delta)$ statistic is typically viewed as a D study statistic, but drawing that distinction in the above discussion seems unnecessarily specific.

In this paper we typically use “G study” to refer to the set of data for an operational measurement procedure, as opposed to the UAO which may contain more facets than those that are identifiable in a G study. Our primary concern is to examine the consequences of having facets in the UAO that are confounded in a particular G study.

In papers that address other features of G theory, it is often more natural to associate the phrase “measurement procedure” with a D study. Here, we generally refer to ”D study” whenever we want to focus attention on parameters such as error variances and coefficients that are based on specific D study sample sizes, with explicit consideration of which facets are fixed and which are random. Note that we do not consider cases in which the G and D study are based on different sets of data.

1.2.3 Replications do Not Eliminate Confounding

In statistics and psychometrics, doubts about the credibility of results for a single study get reduced if sample sizes increase and/or the study is replicated and the results for the two studies are similar. This is *not* true when both studies involve the same type of confounding. Confounding is a *structural* problem that persists with replicated studies and/or larger sample sizes.

A common example of this misunderstanding occurs when an entity claims that its essay scoring procedures are highly reliable based on the fact that correlations are all quite high for a large number of $p \times r$ analyses with two raters, each of which uses a single, diferent prompt. Since every one of these analyses uses a single prompt, the correlation is given by Equation 6 with $n'_r = 1$ (see Brennan, 2001, p. 129). The crucial point to note is that the numerator of Equation 6 involves $\sigma^2(pz)$ for every one of these analyses.[4] This inflates the correlation relative to Equation 8 with $n'_r = 1$, which is for the intended UAO in which prompts are random. In short, “averaging” over multiple analyses, all of which involve a different, single prompt simply propagates the confounding over the multiple analyses. Doing so does *not* eliminate, or even mitigate, the confounding.

1.2.4 Multiple G Studies

In the above discussion of the G study $p \times r$ design with a single fixed level of z, it was noted that it is impossible to disentangle $\sigma^2(p)$ and $\sigma^2(pz)$. That is true if the only available G study is for the $p \times r$ design. This problem can be circumvented, however, if there is an auxillary G study that involves at least two levels of z and that has p crossed with z. Such an auxillary study may have fewer examinees and/or have other limitations, but the study may still help in disentangling $\sigma^2(p)$ and $\sigma^2(pz)$, at least approximately. Gao et al. (2015) discuss the use of multiple G studies to approximate error variance and coefficients under complex measurement conditions.

2 A Single Prompt for Different Prompt Types

Suppose an essay test consists of two types of essay prompts (e.g., narrative and persuasive), and suppose that:

[4] The variance component $\sigma^2(pz)$ reflects the person-prompt interaction, which is almost always quite large.

- the two types of prompts (a) are fixed in the sense that different forms of the assessment would contain the *same* two types; and
- the actual essay prompts (z) are random in the sense that different forms of the assessment would contain *different* prompts.

Structurally, we say that z is nested within a in the UAO). We denote this nesting as $z : a$. Suppose, as well, that all examinees or persons (p) take the same prompts, which are scored by human raters (r) such that:

- different raters are used for each prompt; and
- different raters are used over forms.

This means that raters are random, and they are nested within prompts in the UAO. It follows that $r\colon z\!:\!a$. Assuming all examinees or persons (p) respond to all prompts, we denote the full design as $p \times (r\colon z\!:\!a)$.

G theory can handle the above design in a relatively straightforward manner provided there are at least two conditions (levels or instances) of each of the facets r, z, and a in a G Study. Confusion and complexities arise, however, when there is only a single condition for r and/or z in the G Study. (We assume here that both fixed conditions of a are always present.) Suppose, for example, that in the G study there is only one prompt for each type of prompt. In such a case, we say that prompt z and type of prompt a are "completely confounded," which we denote (z, a). This means that picking a specific level of z involves simultaneously picking a specific level of a, and vice-versa.

The completely confounded notation notation (z, a) needs to be distinguished from:

- $z\!:\!a$, which means z is nested within a; and
- $z|a$ (as in section 1.1), which means that each level of z involves the same single, fixed level of a.

Clearly, $z|a$ involves confounding, but here we generally reserve the term "completely confounded" for (z, a).

We consider the hypothetical essay-test example outlined above from several different perspectives that force us to distinguish carefully between different universes and designs, between which facets are fixed and which are random, and whether or not a D study sample size is one. We begin with a careful consideration of the variance components and related statistics that we would like to know. Subsequently, we consider the results that are available to us in certain operational testing circumstances.

The remaining subsections (2.1–2.6) in this section are somewhat complicated. For some readers, skimming these sections may be sufficient.

Table 1: Parametric Values for Variance Components in the UAO based on the $p \times (r\colon z\,{:}\,a)$ Design; and for D Study Variance Components, Error Variances, and Coefficients with A Fixed, Z Random, R Random, and $n'_z = 1$

G Study Variance Components		D Studies with A Fixed		
		n'_a	2	2
		n'_z	1	1
A Random	A Fixed	n'_r	2	1
$\sigma^2(p) = .25$	$\sigma^2(p\,\vert A) = .29$	$\sigma^2(p\,\vert A)$	.2900	.2900
$\sigma^2(a) = .03$				
$\sigma^2(z\,{:}\,a) = .06$	$\sigma^2(z\,{:}\,a\vert A) = .06$	$\sigma^2(Z\,{:}\,A\vert A)$	.0300	.0300
$\sigma^2(r\colon z\,{:}\,a) = .02$	$\sigma^2(r\colon z\,{:}\,a\vert A) = .02$	$\sigma^2(R\colon Z\,{:}\,A\vert A)$	.0050	.0100
$\sigma^2(p\,a) = .08$				
$\sigma^2(p\,z\,{:}\,a) = .10$	$\sigma^2(p\,z\,{:}\,a\vert A) = .10$	$\sigma^2(p\,Z\,{:}\,A\vert A)$	.0500	.0500
$\sigma^2(p\,r\colon z\,{:}\,a) = .12$	$\sigma^2(p\,r\colon z\,{:}\,a\vert A) = .12$	$\sigma^2(p\,R\colon Z\,{:}\,A\vert A)$	.0300	.0600
		$\sigma^2(\tau)$	.2900	.2900
		$\sigma^2(\delta)$	.0800	.1100
		$\sigma^2(\Delta)$	.1150	.1500
		$\boldsymbol{E}\rho^2$	.7838	.7250
		Φ	.7160	.6591

Note. $\sigma^2(a\vert A)$ is not listed in the second column because it is a quadratic form. $\sigma^2(A\vert A)$ disappears in the last two columns which involve the average over the two fixed levels of a; i.e., $\sigma^2(A\vert A) = 0$.

2.1 UAO, G Study, and Projected D study Results

The second column of Table 1 provides hypothetical variance components for the UAO. The results in the second column are for type-of-prompt being fixed, which is denoted by appending "$\vert A$" to the notation for each variance component. Often, in practice, the results in the second column are obtained using the random effects variance components in the first column, which assume that type of prompt is random. For example,

$$\sigma^2(p\vert A) = \sigma^2(p) + \frac{\sigma^2(pa)}{2}, \tag{9}$$

where variance components to the right of the equal sign are for the random model in the first column.[5] Equation 9 is a special case of the general procedure discussed by Brennan (2001, pp. 120–124) for obtaining mixed-model variance components from random model variance components.

Although the variance components in the second column are hypothetical, they are not entirely unreasonable. For example, $\sigma^2(z\,{:}\,a) = .06$ is three times

[5] As noted earlier, the notation in this paper does not distinguish between parameters and estimates; doing so seems unnecessary here, and would likely add more confusion than clarity for most readers.

larger than the variance component $\sigma^2(r\colon z\!:\!a) = .02$, which means that prompts are considerably more variable than raters. This often happens for testing programs that have good rubrics and well-trained raters. It is generally much easier to get raters to behave similarly than it is to get passages to be equally difficult.

The third column provides the notational conventions for denoting D study variance components and other results; uppercase letters denote mean scores. At the bottom of the column, $\sigma^2(\tau)$ is universe score variance, $\sigma^2(\delta)$ is relative error variance, $\sigma^2(\Delta)$ is absolute error variance, $\boldsymbol{E}\rho^2$ denotes a generalizability coefficient, and Φ denotes a phi coefficient. Formulas for obtaining these results are provided by Brennan (2001).

The fourth column provides "projected" D study results for the universe of generalization (UG) in which A is fixed with a sample size of $n'_a = 2$, Z is random with a sample size of $n'_z = 1$, and R is random with a sample size of $n'_r = 2$. It is important to understand that these results are *not* based on a data collection design in which there is only one prompt ($n'_z = 1$) for each of the two levels of the a facet. Rather, the results in the fourth column are based on the UAO variance components in the second column, which we assume are known or could be estimated using a G study with at least two levels of all facets. If all the results in columns 1 or 2 were actually available, then confounded effects would not be problematic.

Because the results in the fourth column are *not* based on an actual D study with $n'_z = 1$, we refer to them above as "projected" D study results. They are the results of interest for a D study with $n'_z = 1$. As discussed below and illustrated in the next two subsections, these results *cannot* be obtained using an operational assessment with $n'_z = 1$.

The statements in the above paragraph may appear contradictory or inconsistent. After all, how can "projected" D study results be better than results for an operational assessment? The explanation for this apparent contradiction rests on recognizing the following facts:

1. the prompt facet is intended to be random, while the prompt-type facet is intended to be fixed;

2. when $n'_z = 1$, prompt and prompt type are completely confounded;

3. any analysis of an operational assessment in which (2) holds must treat the prompt and prompt-type facets as both fixed or both random, which contradicts (1);

4. the "projected" D study results in the fourth column are entirely consistent with both (1) and (2).

If available data contain results for only one prompt for each prompt type, then the effects for the two facets cannot be distinguished in any manipulation of the data. Still we must decide whether to treat the confounded effect as random or fixed, and the choice makes a difference that can be substantial.

Table 2: Variance Components for $p \times (r\!:\!(z,a))$ design, and D Study Results for $p \times (R\!:\!(Z,A))$ Design with R Random and (Z,A) Random

	D Studies for Three Random Facets		
Variance Components for	$n'_{(z,a)}$	2	2
Single Conditions of Facets	n'_r	2	1
$\sigma^2(p) = .25$	$\sigma^2(p)$	.2500	.2500
$\sigma^2(z,a) = .09$	$\sigma^2(Z,A)$	.0450	.0450
$\sigma^2(r\!:\!(z,a)) = .02$	$\sigma^2(R\!:\!(Z,A))$	.0050	.0100
$\sigma^2(p\,(z,a)) = .18$	$\sigma^2(p\,(Z,A))$	.0900	.0900
$\sigma^2(p\,r\!:\!(z,a)) = .12$	$\sigma^2(p\,R\!:\!(Z,A))$	.0300	.0600
	$\sigma^2(\tau)$	.2500	.2500
	$\sigma^2(\delta)$	.1200	.1500
	$\sigma^2(\Delta)$	.1700	.2050
	$\boldsymbol{E}\rho^2$	.6757	.6250
	Φ	.5952	.5495

2.2 D Studies with (Z,A) Random

Table 2 provides results for the D study $p \times (R\!:\!(Z,A))$ design when R is random and (Z,A) is treated as random. The variance components in the first column of Table 2 can be obtained easily from the variance components in the first column of Table 1. Specifically, $\sigma^2(p) = .25$ is the same in both columns, and

$$\begin{aligned}
\sigma^2(z,a) &= \sigma^2(a) + \sigma^2(z\!:\!a) \;=\; .03 + .06 \;=\; .09 \\
\sigma^2(r\!:\!(z,a)) &= \sigma^2(r\!:\, z\!:\!a) \;=\; .02 \qquad (10) \\
\sigma^2(p\,(z,a)) &= \sigma^2(pa) + \sigma^2(p\,z\!:\!a) \;=\; .08 + .10 \;=\; .18 \\
\sigma^2(p\,r\!:\!(z,a)) &= \sigma^2(p\,r\!:\, z\!:\!a) \;=\; .12 \qquad (11)
\end{aligned}$$

The sum of the variance components in column 1 of Tables 1 and 2 is the same, namely .66, but the seven variance components in Table 1 collapse to five in Table 2. The sum is the same because all facets are random in both tables.[6]

Of course, if the available data from an operational administration of the assessment involve only one level of z, not all of the variance components in Table 1 would be known (or estimable). That does not mean, however, that these unknown variance components in Table 1 do not exist; they are merely invisible to the investigator who analyzes data from the operational administration. Here, we are simply assuming that the variance components in Table 1 are known, and under this assumption we would expect to obtain the numerical results in Table 2.

[6] In Equation 10, $\sigma^2(r\!:\!(z,a)) = \sigma^2(r\!:\, z\!:\!a)$ because with $n_z = 1$ there is no structural (i.e., design) difference between (z,a) and $z\!:\!a$; a similar statement holds for Equation 11.

The third column in Table 2 provides D study results for $n'_r = 2$ and $n'_{(z,a)} = 2$, where $n'_{(z,a)} = 2$ means that we have two levels of the (prompt, prompt-type) confounded effect.[7] Note, in particular, that the results in the third column of Table 2 are for (Z, A) random. These results can be compared to the values in the fourth column of Table 1 where Z is random and A is fixed. It is clear that

- universe score variance is smaller in Table 2,
- error variances are larger in Table 2, and
- coefficients are substantially smaller in Table 2.

All of these results are directly attributable to the fact that A is treated as random in Table 2 when it should be treated as fixed. Consequently, in Table 2, universe score variance and coefficients are biased downward, and error variances are biased upward. For those who wish a more detained explanation, the following discussion may help.

Universe score variance is biased downward because, from Equation 9, $\sigma^2(\tau)$ with (Z, A) random is

$$\sigma^2(p) = \sigma^2(p\,|A) - \frac{\sigma^2(pa)}{2} = \sigma^2(p|A) - \sigma^2(pA).$$

That is, $\sigma^2(pA)$ gets subtracted from $\sigma^2(p\,|A)$. In addition, $\sigma^2(pA)$ gets added to $\sigma^2(\delta)$ and $\sigma^2(\Delta)$, which causes these error variances to get larger. $\sigma^2(\Delta)$ also gets larger by the addition of $\sigma^2(a)/2$.

2.3 D Studies with (Z, A) Fixed

Table 3 provides results when R is random and (Z, A) is treated as fixed. The variance components in the first column of Table 3 can be obtained from the variance components in the second column of Table 1. Specifically,

$$\begin{aligned} \sigma^2(r\!:\!(z,a)|(Z,A)) &= \sigma^2(r\!:\,z\!:\!a|A) = .02, && (12)\\ \sigma^2(p\,r\!:\!(z,a)|(Z,A)) &= \sigma^2(p\,r\!:\,z\!:\!a|A) = .12, && (13) \end{aligned}$$

and it can be shown that

$$\sigma^2(p\,|(Z,A)) = \sigma^2(p\,|A) + \frac{\sigma^2(p\,z\!:\!a|A)}{2} = .29 + \frac{.10}{2} = .34. \qquad (14)$$

Alternatively,

$$\sigma^2(p\,|(Z,A)) = \sigma^2(p) + \frac{\sigma^2(pa)}{2} + \frac{\sigma^2(p\,z\!:\!a)}{2} = .25 + \frac{.08}{2} + \frac{.10}{2} = .34.$$

[7] The fourth column in Table 2 is considered later in Section 2.6.

Table 3: Variance Components for $p \times (r\!:\!(z,a))$ design, and D Study Results for $p \times (R\!:\!(Z,A))$ Design with R Random and (Z,A) Fixed

	D Studies for One Random Facet		
Variance Components for	$n'_{(z,a)}$	2	2
Single Conditions of Facets	n'_r	2	1
$\sigma^2(p\,\|(Z,A)) = .34$	$\sigma^2(p\,\|(Z,A))$	.3400	.3400
$\sigma^2(r\!:\!(z,a)\|(Z,A)) = .02$	$\sigma^2(R\!:\!(Z,A)\|(Z,A))$	.0050	.0100
$\sigma^2(p\,r\!:\!(z,a)\|(Z,A)) = .12$	$\sigma^2(p\,R\!:\!(Z,A)\|(Z,A))$	.0300	.0600
	$\sigma^2(\tau)$	.3400	.3400
	$\sigma^2(\delta)$	.0300	.0600
	$\sigma^2(\Delta)$	.0350	.0700
	$\boldsymbol{E}\rho^2$	.9189	.8500
	Φ	.9067	.8293

The last equation results from an application of a the general procedure discussed by Brennan (2001, pp. 120–124) for obtaining mixed-model variance components from random model variance components.[8]

The third column in Table 3 provides the D study results for $n'_r = 2$ and $n'_{(z,a)} = 2$, with (Z,A) fixed. These results can be compared to the values in the fourth column of Table 1 with Z random and A fixed. It is clear that

- universe score variance is larger in Table 3,
- error variances are smaller in Table 3, and
- coefficients are substantially larger in Table 3.

All of these results are directly attributable to the fact that Z is treated as fixed in Table 3 when it should be treated as random. Consequently, in Table 3, universe score variance and coefficients are biased upward, and error variances are biased downward. For those who wish a more detained explanation, the following discussion may help.

Universe score variance for (Z,A) fixed is larger than for A fixed and Z random. This occurs because, with (Z,A) fixed, as indicated in Equation 14, $\sigma^2(p\,z : a|A)/2$ gets added to $\sigma^2(p\,|A)$, which is universe score variance for A fixed. Note also that $\sigma^2(p\,z : a|A)/2$ gets subtracted from $\sigma^2(\delta)$ and $\sigma^2(\Delta)$,

[8] In Equation 12, $\sigma^2(r\!:\!(z,a)|(Z,A)) = \sigma^2(r\!: z\!:\!a|A)$ because the variance component is for raters, and it does not matter what facets or types of facets the raters are nested within. A similar statement holds for person-rater combinations in Equation 13.

which causes these error variances to get smaller. The error variance $\sigma^2(\Delta)$ also gets smaller by the subtraction of $\sigma^2(z\!:\!a|A)/2$.

Tables 1–3 provide results for $n'_r = 1$ and $n'_r = 2$, which are the most common sample sizes in most performance assessments. That is, typically there is only one or two ratings of the performance of any single person on any single prompt.

2.4 Results for $n'_r = 2$

The bulleted results in Sections 2.2 and 2.3 clearly indicate that the (Z, A) random results and the (Z, A) fixed results bracket the parameter-of-interest values in the fourth column of Table 1. Specifically,

$n'_r = 2$	(Z,A) Random		A Fixed		(Z,A) Fixed
$\sigma^2(\tau)$	.2500	$<$	.2900	$<$	.3400
$\sigma^2(\delta)$	.1200	$>$	.0800	$>$	.0300
$\sigma^2(\Delta)$	.1700	$>$	.1150	$>$	.0350
$\boldsymbol{E}\rho^2$	.6757	$<$	.7838	$<$	.9189
Φ	.5952	$<$	.7160	$<$	.9067

One implication of the above results is that, if the only available data are for a study in which the prompt and prompt-type facets are completely confounded, then it is appropriate (and indeed desirable) to perform analyses with (Z, A) random *and* with (Z, A) fixed.[9] Then, an interval for each parameter-of-interest (i.e., a lower and upper bound) can be reported.

2.5 Results for $n'_r = 1$

The last columns in Tables 1–3 provide the same types of results discussed previously, with the one difference being that $n'_r = 1$. That is, these results are for a measurement procedure in which there are two fixed types of prompts, with a single random prompt for each type, and with the responses to each prompt rated by a single random rater. Again, we see that the parameter-of-interest results in Tables 1 are bracketed by the results in Tables 2 and 3. Specifically,

$n'_r = 1$	(Z,A) Random		A Fixed		(Z,A) Fixed
$\sigma^2(\tau)$	.2500	$<$	.2900	$<$	.3400
$\sigma^2(\delta)$	.1500	$>$	.1100	$>$	.0600
$\sigma^2(\Delta)$	.2050	$>$	.1500	$>$	.0700
$\boldsymbol{E}\rho^2$	.6250	$<$	.7250	$<$	.8500
Φ	.5495	$<$	.6591	$<$	.8293

[9] The expected observed score variance, $\sigma^2(\tau) + \sigma^2(\delta)$, is .37 for all three analyses, because expected observed score variance is unaffected by which facets are fixed and which are random.

As must be the case, $\boldsymbol{E}\rho^2$, and Φ for $n'_r = 1$ are smaller than for $n'_r = 2$, and error variances are larger. Otherwise, however, the conclusions here mirror those in the previous subsection.

Note that the widths of the intervals for $\sigma^2(\tau)$, $\sigma^2(\delta)$, and $\sigma^2(\Delta)$ (.09, .09, and .135, respectively) are the same for $n'_r = 1$ and $n'_r = 2$.[10] This occurs because, for each of these statistics, going from (Z, A) random to (Z, A) fixed leads to the addition or subtraction of a constant that is unrelated to the sample size for raters. For $\boldsymbol{E}\rho^2$ and Φ, the widths differ solely because the error variances are larger for $n'_r = 1$ than for $n'_r = 2$.

When both $n'_r = 1$ and $n'_z = 1$, it is obvious that raters, prompts, and prompt types are all confounded, which we can denote (r, z, a), or (R, Z, A). For the example considered here, this triple confounding implies that there are only two observations for each person, since $n_a = n'_a = 2$ and the other two facets have a sample size of 1. Under these circumstances, it is not sensible to treat (r, z, a) as fixed, because then there are no random facets, which means that error variances are all 0 and coefficients are all 1. So, in this case, the only analysis that can be done treats (r, z, a) as random, which is discussed next.

2.6 Coefficient Alpha and $(\boldsymbol{R}, \boldsymbol{Z}, \boldsymbol{A})$

If (r, z, a) is viewed as random, then the analysis will proceed as if there is one random facet that is crossed with persons. This is the single-facet crossed design discussed extensively by Brennan (2001, chap. 2). The resulting generalizability coefficient is identical to Coefficient alpha.[11] For our hypothetical example, the value of Coefficient alpha that we would expect to obtain is the generalizability coefficient in the last column of Table 2, namely, .6250, which is less than the parameter value of .7250 in the last column of Table 1 with the prompt-type facet fixed.

The fact that the Coefficicent alpha value of .6250 is less that the parameter value of .7250 has nothing to do with the mathematical proof in Lord and Novick (1968, pp. 88–90) that alpha is a lower limit to reliability. In effect, their proof is based on the assumption that Coefficient alpha and the parameter involve the same sources of error, all of which are confounded in a single "clump" of random error. For the example considered here, the value of alpha that would be expected based on the (D study) data is a *lower* limit because this value (.6250) implicitly involves three random facets, whereas the parameter value (.7250) involves only two random facets; i.e., the parameter uses a more restrictive definition of error than does alpha.

Note that $\boldsymbol{E}\rho^2 = .6757$ in the third column of Table 2 does *not* have a Coefficient-alpha type of interpretation. In that column, $\boldsymbol{E}\rho^2$ is based on four observations per examinee: two raters who evaluate one (prompt, prompt type) and a different two raters who evaluate the second (prompt, prompt type). That

[10] The fact that the widths of the $\sigma^2(\tau)$ and $\sigma^2(\delta)$ intervals are the same (.09) is purely an artifact of these particular, hypothetical data.

[11] Brennan (2001, p. 128) provides another perspective on Coefficient alpha and confounding due to the fact that data for computing Coefficient alpha usually come from a single occasion.

is, the design in Table 2 is $p \times (R\!:\!(Z,A))$; and with $n'_r = 2$, the rater facet is not confounded with (prompt, prompt type). It follows that the relative error variance in $\boldsymbol{E}\rho^2 = .6757$ is

$$\sigma^2(\delta) = \frac{\sigma^2(p\,(z,a))}{2} + \frac{\sigma^2(p\,r\!:\!(z,a))}{4} = \frac{.18}{2} + \frac{.12}{4} = .9 + .3 = .12.$$

By contrast, if the data were analyzed as if each of the four observations for each person were independent (the assumption for Coefficient alpha), then the error variance would be

$$\frac{\sigma^2(p\,(z,a)) + \sigma^2(p\,r\!:\!(z,a))}{4} = \frac{.18 + .12}{4} = .0750,$$

and Coefficient alpha would be $.25./(.25 + .0750) = .7692$. That is, in this case, Coefficient alpha would be an *upper* limit to reliability.

In short, this section illustrates that whether Coefficient alpha is a lower limit or an upper limit to reliability (in the sense of a generalizability coefficient) depends on which facets are fixed, which facets are random, the D study design, and the D study sample sizes. Whenever conditions of a random and fixed facet are confounded in the data for a measurement procedure, Coefficient alpha is an inappropriate estimate of reliability.

3 A Single Rater (e.g., Automated Scoring Engine)

Thus far in this paper, we have considered a UAO and designs in which raters are nested within another facet. Now let us consider a UAO in which raters are crossed with other facets and with persons. When the number of persons is very large, it is highly unlikely that a G study design with human raters would faithfully mirror this UAO, because every rater would have to evaluate every response or performance for every examinee.

Currently, however, there is considerable interest in the use of automated scoring engines to rate examinee responses. Since any particular automated scoring engine employs a particular algorithm, that algorithm functions like a rater that is the same for all examinees and other facets. Of course, there are different automated scoring engines, each of which uses a different algorithm. Unless otherwise noted, in this section we assume the UAO contains a facet consisting of numerous possible automated scoring engines and, for consistency with the previous sections, we refer to this facet as the rater facet.

3.1 Example

For the $p \times r \times (z\!:\!a)$ design, Table 4 extends the hypothetical example originally introduced in Table 1. With respect to column one in both tables, the reader can verify that

$$\sigma^2(r\!:\, z\!:\!a) = \sigma^2(r) + \sigma^2(ra) + \sigma^2(rz\!:\!a) = .010 + .006 + .004 = .020,$$

Table 4: Parametric Values for Variance Components in the UAO for the $p \times r \times (z\!:\!a)$ Design, and for D Study Variance Components, Error Variances, and Coefficients with A Fixed, Z Random, R Random, and $n'_z = 1$

		D Studies with A Fixed		
Variance Components in UAO		n'_a	2	2
		n'_z	1	1
A Random	A Fixed	n'_r	2	1
$\sigma^2(p) = .250$	$\sigma^2(p\,\vert A) = .290$	$\sigma^2(p\,\vert A)$	.2900	.2900
$\sigma^2(r) = .010$	$\sigma^2(r\,\vert A) = .013$	$\sigma^2(R\,\vert A)$	.0065	.0130
$\sigma^2(a) = .030$				
$\sigma^2(z\!:\!a) = .060$	$\sigma^2(z\!:\!a\vert A) = .060$	$\sigma^2(Z\!:\!A\vert A)$	.0300	.0300
$\sigma^2(p\,r) = .040$	$\sigma^2(p\,r\vert A) = .050$	$\sigma^2(p\,R\vert A)$	.0250	.0500
$\sigma^2(p\,a) = .080$				
$\sigma^2(p\,z\!:\!a) = .100$	$\sigma^2(p\,z\!:\!a\vert A) = .100$	$\sigma^2(p\,Z\!:\!A\vert A)$	.0500	.0500
$\sigma^2(r\,a) = .006$				
$\sigma^2(rz\!:\!a) = .004$	$\sigma^2(rz\!:\!a\vert A) = .004$	$\sigma^2(RZ\!:\!A\vert A)$	.0010	.0020
$\sigma^2(pra) = .020$				
$\sigma^2(prz\!:\!a) = .060$	$\sigma^2(prz\!:\!a\vert A) = .060$	$\sigma^2(pRZ\!:\!A\vert A)$	.0150	. 0300
		$\sigma^2(\tau)$	.2900	.2900
		$\sigma^2(\delta)$	.0900	.1300
		$\sigma^2(\Delta)$	.1275	.1750
		$\boldsymbol{E}\rho^2$	.7632	.6905
		Φ	.6946	.6237

Note. $\sigma^2(a|A)$ is not listed in the second column because it is a quadratic form. $\sigma^2(A|A)$ disappears in the last two columns which involve the average over the two fixed levels of a; i.e., $\sigma^2(A|A) = 0$.

and

$$\sigma^2(pr\colon z\!:\!a) = \sigma^2(pr) + \sigma^2(pra) + \sigma^2(prz\!:\!a) = .040 + .020 + .060 = .120$$

where variance components to the right of the equal sign (for the $p \times r \times (z : a)$ design) provide a decomposition of the variance components to the left of the equal sign (for the $p \times (r\colon z\!:\!a)$ design). The numerical values for the variance components in Table 4 are hypothetical, but they bear some similarities with those in Table 1.

The second column in Table 4 provides variance components for the UAO in which A is fixed. Note that:

$$\sigma^2(p|A) = \sigma^2(p) + \frac{\sigma^2(pa)}{2},$$

$$\sigma^2(r|A) = \sigma^2(r) + \frac{\sigma^2(ra)}{2},$$

and

$$\sigma^2(pr|A) = \sigma^2(pr) + \frac{\sigma^2(pra)}{2}.$$

The last two columns provide D study results for the sample sizes considered throughout this paper. Note that the next-to-the-last column provides results for $n'_r = 2$ for consistency with Table 1 (and other tables in Section 1), but our principal focus here is results for $n'_r = 1$ in the last column.

Of particular importance is the fact that the error variances are larger and the coefficients are smaller in Table 4, where raters are crossed with (z, a), than in Table 1, where raters are nested within (z, a). A somewhat heuristic explanation for this is that in Table 4 there are only n'_r raters involved in the design, whereas in Table 1 there are $2n'_r$ raters involved (e.g., when $n'_r = 1$ there is one rater for (z_1, a_1) and a different rater for (z_2, a_2)). In general, for a given value of n'_r, $\sigma^2(\delta)$ in Table 4 is larger than $\sigma^2(\delta)$ in Table 1 by $\sigma^2(pr)/(2n'_r)$, and $\sigma^2(\Delta)$ in Table 4 is larger that $\sigma^2(\Delta)$ in Table 1 by $(\sigma^2(r) + \sigma^2(pr))/(2n'_r)$. (See the Appendix for further, more mathematical explanations.) So, all other things being equal, the crossed design has larger error variances (and smaller coefficients) than the nested design.

3.1.1 R Random and (Z, A) Treated as Random

Suppose the only available data are from the operational administration of a performance assessment that uses the $p \times r \times (z\!:\!a)$ design with the prompt and prompt-type facets being completely confounded. Table 5 provides results for the D study $p \times R \times (Z\!:\!A)$ design when R is random and (Z, A) is treated as random.

The variance components in the first column of Table 5 can be obtained from the variance components in the first column of Table 4. Specifically, $\sigma^2(p)$, $\sigma^2(r)$, and $\sigma^2(pr)$ are unchanged, and

$$\begin{aligned}
\sigma^2(z,a) &= \sigma^2(a) + \sigma^2(z\!:\!a) \;=\; .030 + .060 \;=\; .090 \\
\sigma^2(p\,(z,a)) &= \sigma^2(pa) + \sigma^2(p\,z\!:\!a) \;=\; .080 + .100 \;=\; .180 \\
\sigma^2(r(z,a)) &= \sigma^2(ra) + \sigma^2(rz\!:\!a) \;=\; .006 + .004 \;=\; .010 \qquad (15) \\
\sigma^2(pr(z,a)) &= \sigma^2(pra) + \sigma^2(prz\!:\!a) \;=\; .020 + .060 \;=\; .080, \qquad (16)
\end{aligned}$$

where variance components to the left of the equal sign are for Table 5. The sum of the variance components in column 1 of Tables 4 and 5 is the same, namely .66, but the 11 variance components in Table 4 collapse to seven in Table 5. The sum is the same because all facets are random in both tables.[12]

The variance components in the first column of Table 5 are related to the variance components in the first column of Table 2. Specifically, $\sigma^2(p)$, $\sigma^2(z\!:\!a)$,

[12] In Equation 15, $\sigma^2(rz\!:\!a) = \sigma^2(r(z,a))$ because: (a) $\sigma^2(rz\!:\!a)$ is an abbreviated notation for $\sigma^2(r(z:a))$, and (b) with $n_z = 1$ there is no structural (i.e., design) difference between $r(z\!:\!a)$ and $r(z,a)$; similarly, in Equation 16 $\sigma^2(prz\!:\!a) = \sigma^2(pr(z,a))$.

Table 5: Variance Components for $p \times r \times (z\!:\!a)$ design, and D Study Results for the $p \times R \times (Z\!:\!A)$ Design with R Random and (Z, A) Random

	D Studies for Three Random Facets		
Variance Components for	$n'_{(z,a)}$	2	2
Single Conditions of Facets	n'_r	2	1
$\sigma^2(p) = .250$	$\sigma^2(p)$	.2500	.2500
$\sigma^2(r) = .010$	$\sigma^2(R)$	,0050	.0100
$\sigma^2(z,a) = .090$	$\sigma^2(Z,A)$	.0450	.0450
$\sigma^2(pr) = .040$	$\sigma^2(pR)$	.0200	.0400
$\sigma^2(p\,(z,a)) = .180$	$\sigma^2(p\,(Z,A))$	.0900	.0900
$\sigma^2(r(z,a)) = .010$	$\sigma^2(R(Z,A))$	.0025	.0050
$\sigma^2(pr(z,a)) = .080$	$\sigma^2(pR(Z,A))$	.0200	.0400
	$\sigma^2(\tau)$	.2500	.2500
	$\sigma^2(\delta)$	.1300	.1700
	$\sigma^2(\Delta)$	.1825	.2300
	$\boldsymbol{E}\rho^2$	.6579	.5952
	Φ	.5780	.5208

Table 6: Variance Components for $p \times r \times (z, a)$ design, and D Study Results for $p \times R \times (Z, A))$ Design with R Random and (Z, A) Fixed

	D Studies for One Random Facet		
Variance Components for	$n'_{(z,a)}$	2	2
Single Conditions of Facets	n'_r	2	1
$\sigma^2(p\,\vert(Z,A)) = .340$	$\sigma^2(p\,\vert(Z,A))$	.3400	.3400
$\sigma^2(r\vert(Z,A)) = .015$	$\sigma^2(R\!:\!(Z,A)\vert(Z,A))$	.0075	.0150
$\sigma^2(pr\vert(Z,A)) = .080$	$\sigma^2(p\,R\!:\!(Z,A)\vert(Z,A))$	.0400	.0800
	$\sigma^2(\tau)$	.3400	.3400
	$\sigma^2(\delta)$	.0400	.0800
	$\sigma^2(\Delta)$	.0475	.0950
	$\boldsymbol{E}\rho^2$	.8947	.8092
	Φ	.8774	.7816

and $\sigma^2(p\,(z,a))$ are unchanged. Also,

$$\begin{aligned} \sigma^2(r\!:\!(z,a)) &= \sigma^2(r) + \sigma^2(r(z,a)) \;=\; .010 + .010 \;=\; .020 \quad \text{and} \\ \sigma^2(p\,r\!:\!(z,a))) &= \sigma^2(pr) + \sigma^2(pr(z,a)) \;=\; .040 + .080 \;=\; .120, \end{aligned}$$

where variance components to the left of the equal signs are for Table 2. These equations could be used to obtain results for the nested $p \times (r\!: z\!:\!a)$ design from the crossed $p \times r \times (z\!:\!a)$ design.

3.1.2 R Random and (Z, A) Treated as Fixed

Table 6 provides results for the D study $p \times R \times (Z\!:\!A)$ design when R is random and (Z, A) is treated as fixed. The variance components in the first column of Table 6 can be obtained from the variance components in the first column of Table 5. Specifically,

$$\sigma^2(p\,|(Z,A)) = \sigma^2(p) + \frac{\sigma^2(p\,(z,a))}{2},$$

$$\sigma^2(r|(Z,A)) = \sigma^2(r) + \frac{\sigma^2(r(z,a))}{2},$$

and

$$\sigma^2(pr|(Z,A)) = \sigma^2(pr) + \frac{\sigma^2(pr(z,a))}{2},$$

where variance components to the right of the equal sign are in Table 6.

3.1.3 Results for $n'_r = 1$

The last columns of Tables 4, 5, and 6 provide results for the case of $n'_r = 1$, which means that the single rater (automated scoring engine) facet is confounded with persons and the other facets. These results are summarized below:

$n'_r = 1$	(Z, A) Random		A Fixed		(Z, A) Fixed
$\sigma^2(\tau)$	.2500	$<$	.2900	$<$	.3400
$\sigma^2(\delta)$	.1700	$>$	.1300	$>$	.0800
$\sigma^2(\Delta)$	.2300	$>$	.1750	$>$	.0950
$\boldsymbol{E}\rho^2$	.5952	$<$	.6905	$<$	.8092
Φ	.5208	$<$	.6237	$<$	.7816

Comparing these results with those in Section 2.5, it is evident that for a single rater nested within facets, coefficients are larger and error variances are smaller, relative to what they are for a single rater (automated scoring engine) crossed with all persons and other facets. This is true whether or not the prompt facet (z) and the prompt-type facet (a) are confounded. The numerical results discussed above are for hypothetical data, but the conclusions hold in general provided the UAO's contain the same facets and the population is the same.

3.2 Other Issues

The previous section by no means exhausts the conceptual, design, or computational complexities involved in estimating error variances and reliability-like coefficients when an ASE is used. A few additional complexities are briefly discussed next.

3.2.1 Single Random vs. Fixed ASE

The discussion in Section 3, as it relates to $n'_r = 1$, is relevant for a single *random* ASE, but estimating results presumes that a G study is conducted that employs at least two ASEs. If a G study includes only one ASE, then the rater (ASE) facet will be confounded with all other facets as well as the person "facet." This means that any D study results will treat the ASE as a single level of a fixed facet, whether or not that is the intent of the investigator.

For the owner of an ASE whose only concern is that particular ASE, the above paragraph may be interpreted as meaning that Section 3 is irrelevant since it relates to a single *random* ASE. An ASE user, however, may have a legitimate concern about how any single ASE might perform.

3.2.2 Training ASEs

When an ASE is used, it must be "trained" using a set (s) of "papers" that are representative of the responses by the examinees. In the absence of evidence to the contrary, there is no guarantee that the ASE will function the same way if a different representative set of papers were used and/or the training papers were for different subpopulations of examinees (e.g., non-minorities and minorities). So, results for the particular ASE are confounded with the set of papers chosen for training.

Consequently, it seems clear that results for any specific ASE should be examined based on:

1. different sets of papers sampled from the the same population of examinees, and

2. different sets of papers sampled from different subpopulations of examinees.

Differences in results for #1 reflect random error attributable to sampling of papers. Differences in results for #2 reflect one type of differential ASE functioning relevant to subpopulations.

3.2.3 Comparing Multiple ASEs

There are at least two general approaches that might be taken to comparing results for k fixed ASEs. First, conduct k separate univariate G theory (UGT) analyses. Second, conduct a single multivariage G theory (MGT) analysis

A simple example of the UGT approach is to apply the basic methodology in Section 1.1 to each of the k ASEs by interchanging the roles of raters and prompts. Recall that Section 1.1 considers a $p \times r$ design with a single prompt. Here, however, we want an analysis for a $p \times z$ design with a single ASE (which plays the role of a rater). Equations for doing so are obtained by interchanging z and a everywhere in Section 1.1. Obviously, it must be true that $n_z \geq 2$, and the comparisons are sensible only if the persons and prompts are the same for all k ASEs.

A more general (and informative) MGT approach would use the $p^\bullet \times z^\bullet$ design, as discussed in Brennan (2001, chap. 9). An analysis using this design yields three $k \times k$ symmetric variance-covariance matrices ($\mathbf{\Sigma}_p$, $\mathbf{\Sigma}_z$, and $\mathbf{\Sigma}_{pz}$). The diagonal elements are the variance components for the k univariate analyses in the UGT approach. The off-diagonal elements are covariance components for the $k(k-1)/2$ comparisons of pairs of ASEs.

3.2.4 Comparisons Involving Human Raters and an ASE

A very common approach to justifying the use of a particular ASE is to compare the ASE ratings for a single prompt to ratings for human raters, and declare that the ASE is working well if the ASE ratings are similar to those for human raters, in some sense. Presumably, any replicated analysis would use the same ASE but a different set of human raters. (Actually, in most cases, the ASE/rater comparisons are even more complicated because different human raters are often used with different sets of examinees, but we overlook this complexity here.)

A substantial problem with such ASE/rater comparisons is that the ASE is almost always viewed as fixed but the human rater facet is viewed as random. Under these circumstances, any ASE/rater comparisons are at best ambiguous. Stated differently, in a sense, ASE/rater comparisons give the ASE ratings a privileged status in that ASE ratings are not subject to any a priori explicitly represented sources of error.

One way to incorporate potential error in ASE ratings is to conduct a study that includes a facet for different sets of training papers. Then, even for a particular ASE, there will be potentially different ratings for each examinee's response to a prompt. Under these circumstances, there are many specific designs and analyses that might be conducted. (Specifics are outside the intended scope of this paper.) The basic principle, however is clear, namely that a fair comparison of human ratings and ASE ratings requires that ASE ratings incorporate the existence of potential error in such ratings.

4 Concluding Comments

The primary focus of this paper is on confounded effects that arise when,

- for one or more facets, there are many possible conditions in a UAO, but
- a measurement procedure includes only a single condition for one or more of these facets.

We call this “the problem of one.”

Among the important take-home messages from this paper are the following.

1. When there are many conditions for a single facet in a UAO, but only a single condition in the operational measurement procedure, the facet is hidden in the measurement procedure, which leads to understating error variance and overstating coefficients;

2. When a confounded effect in an operational measurement procedure involves a mixture of random and fixed effects, any analysis using the operational data will result in ambiguous results for error variances and coefficients.

3. When (1 or 2) occurs, conducting a G study with at least two conditions for all facets is almost always necessary to resolve confounded-effects ambiguities.

4. When (2) occurs but no such G study is conducted, reporting the end points of a range of possible values for error variances and coefficients seems sensible. The is consistent with the general position that a fuzzy answer to a meaningful, focused question is better than a precise answer to an unimportant or fuzzy question.

Implicit in these take-home messages is the importance of distinguishing between which facets are fixed and which are random in a UAO, which is a hallmark of G theory, but a neglected topic in most other measurement theories. These messages also emphasize that single conditions of facets in a measurement procedure lead to ambiguous results that cannot be resolved by any manipulation of the data. Disentangling these ambiguous results requires additional studies.

This paper does not cover all issues related to confounding. Indeed, as Cronbach et al. (1972) noted decades ago, G theory has a “protean” quality to it. Many, if not most, in-depth generalizability analyses involve different mixes of complex issues, including confounding. Unfortunately, such issues are far too often neglected. In particular, confounded effects almost always lead to reported statistics (e.g, reliabilities and error variances) that are biased, and failure to recognize this fact can easily lead to flawed score interpretations. In a sense, this paper challenges certain aspects of the “conventional wisdom” in psychometrics.

5 References

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

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.

Gao, X., Brennan, R. L., & Guo, F. (2015). *Modeling Measurement Facets and Assessing Generalizability in a Large-Scale Writing Assessment.* GMAC Research Report RR 15-01. Graduate Management Admission Council, P. O. Box 2969, Restin, Virginia 20195.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement*, *6*, 125–160.

6 Appendix: Error Variance Relationships in Tables 1 and 4

Recall that Table 1 is for the case when raters are nested within (z,a), whereas Table 4 is for the case in which raters are crossed with (z,a). For Table 4,

$$\begin{aligned}\sigma^2(\delta) &= \frac{\sigma^2(pr|A)}{n'_r} + \frac{\sigma^2(p\,z\!:\!a|A)}{n'_a n'_z} + \frac{\sigma^2(p\,r\,z\!:\!a|A)}{n'_a n'_z n'_r} \\ &= \frac{\sigma^2(pr) + \sigma^2(pra)/n'_a}{n'_r} + \frac{\sigma^2(p\,z\!:\!a)}{n'_a n'_z} + \frac{\sigma^2(p\,r\,z\!:\!a)}{n'_a n'_z n'_r} \\ &= \frac{\sigma^2(pr)}{n'_r} + \frac{\sigma^2(pra)}{n'_a n'_r} + \frac{\sigma^2(p\,z\!:\!a)}{n'_a n'_z} + \frac{\sigma^2(p\,r\,z\!:\!a)}{n'_a n'_z n'_r}, \qquad (17)\end{aligned}$$

where the last two equalities are expressed in terms of the UAO random effect variance components in Table 4. For Table 1,

$$\begin{aligned}\sigma^2(\delta) &= \frac{\sigma^2(p\,z\!:\!a|A)}{n'_a n'_z} + \frac{\sigma^2(p\,r\!:\,z\!:\!a|A)}{n'_a n'_z n'_r} \\ &= \frac{\sigma^2(p\,z\!:\!a)}{n'_a n'_z} + \frac{\sigma^2(pr) + \sigma^2(pra) + \sigma^2(p\,r\,z\!:\!a)}{n'_a n'_z n'_r} \\ &= \frac{\sigma^2(p\,z\!:\!a)}{n'_a n'_z} + \frac{\sigma^2(pr)}{n'_a n'_z n'_r} + \frac{\sigma^2(pra)}{n'_a n'_z n'_r} + \frac{\sigma^2(p\,r\,z\!:\!a)}{n'_a n'_z n'_r}, \qquad (18)\end{aligned}$$

where the last two equalities are expressed in terms of the UAO random effects variance components in Table 4.

Subtracting Equation 18 from Equation 17 gives

$$\left[\frac{\sigma^2(pr)}{n'_r} + \frac{\sigma^2(pra)}{n'_a n'_r}\right] - \left[\frac{\sigma^2(pr)}{n'_a n'_z n'_r} + \frac{\sigma^2(pra)}{n'_a n'_z n'_r}\right] = \frac{\sigma^2(pr)}{n'_r}\left(1 - \frac{1}{n'_a}\right) + \frac{\sigma^2(pra)}{n'_a n'_r}\left(1 - \frac{1}{n'_z}\right),$$

which is $\sigma^2(pr)/(2n'_r)$ when $n'_z = 1$ and $n'_a = 2$, as is the case in the body of this paper.

A similar approach shows that $\sigma^2(\Delta)$ in Table 4 is larger that $\sigma^2(\Delta)$ in Table 1 by $(\sigma^2(r) + \sigma^2(pr))/(2n'_r)$.