



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Commentary for special issue of journal of memory and language: Generalizability theory analysis for psycholinguistic applications

Aurelio José Figueredo*, Sally Olderbak

Ethology and Evolutionary Psychology, Department of Psychology, University of Arizona, P.O. Box 210068, Tucson, AZ 85721-0068, USA

ARTICLE INFO

Article history:

Received 7 March 2008

revision received 27 July 2008

Available online 10 September 2008

Keywords:

Generalizability theory

Quasi-F-ratios

Psycholinguistics

Random effects models

Null hypothesis significance testing

ABSTRACT

We propose that the continuing controversies over the use of quasi-*F*-ratios in psycholinguistic research might be circumvented, if not resolved, by the judicious application of Generalizability Theory (GT) analyses. We argue that GT is a logical extension of the basic rationale behind repeated measures Analysis of Variance (ANOVA) and the variance components model upon which GT is ultimately based and upon which the entire logic of the *F*-ratio (quasi or otherwise) rests. GT is especially useful in psycholinguistics research because it affords one the opportunity to assess generalizability across multiple dimensions within the same model, such as individual subjects as well as varying conditions of prime and target words. We will provide an illustrative example of GT based on Forster's (2007) replication of Davis and Lupker's (2006) study in which they tested the effects of frequency discrepancies in target and prime words across individual subjects under varying combinations of frequency and prime.

© 2008 Elsevier Inc. All rights reserved.

With advance apologies to the statistically more sophisticated in the readership, we approach this matter didactically by briefly reviewing a few of the basic concepts of Analysis of Variance (ANOVA) so as to develop a common symbolic lexicon with which to build a bridge to Generalizability Theory (GT) Analysis. By building on the well-known, we hope to shed some light upon the less familiar and render it more accessible to a wider readership.

The fundamental logic of the *F*-ratio

We start with the basic theoretical formula for the *F*-ratio:

$$F = (\sigma_t^2 + \sigma_e^2) / \sigma_e^2 \quad (1)$$

Because any estimated treatment effect (e.g., variance between experimental groups) contains some “error” (e.g., sampling error in a between-groups design), to test

the “true” estimated treatment effects it is necessary to divide this numerator by an appropriate denominator that is a purer estimate of the so-called “error” (e.g., variance within experimental groups). The whole point of the *F*-ratio was to determine whether the hypothetical variance component representing the “true” treatment effect (σ_t^2), without the ever-present “error” component (σ_e^2), is (or is not) equal to zero. Because, in Sir Ronald Fisher's time, we had no way to directly determine this, we instead constructed the *F*-ratio to test that “null hypothesis”. If it was indeed the case that the hypothetical variance component for treatment equaled zero ($\sigma_t^2 = 0$), then the expected value of the *F*-ratio would be 1.0 because it would reduce to simply σ_e^2 / σ_e^2 .

What is typically meant by an “error” term in ANOVA is more properly referred to as a *random effect*. The levels of a random effect are conceptualized as being randomly selected from a population and are thus in principle exchangeable with another sample from the same population. The goal of using random effects is to be able to generalize the results to that population. In conventional ANOVA, individuals (variously referred to persons, sub-

* Corresponding author. Fax: +1 520 621 9306.

E-mail address: ajf@u.arizona.edu (A.J. Figueredo).

jects, or participants) are typically treated as random effects. In contrast, the levels of a *fixed effect* are selected by the experimenter as those levels to which they would like to be able to generalize. Levels of a fixed effect can sometimes be conceptualized as being exhaustive of the population of levels of interest (e.g., male and female), so exchangeability with other potential levels is not an issue. Consequently, one may only generalize to the selected levels of the fixed effect that were included in the experiment. In conventional ANOVA, experimental treatments are typically treated as fixed effects, although that is by no means necessary. In our discussion, because the experimental treatment may itself be either fixed or random, we will instead refer to the source of variance being tested as the “focal” effect. This is important in Psycholinguistic research because we often use samples of experimental stimuli, such as words, that are drawn from a larger population to which we explicitly seek to generalize.

In practice, however, the *computational* formulae for *F*-ratios were based on observed, not hypothetical variances. Furthermore, the nature of the “error” terms depended on the relationship of this confounding variance to the “treatment” effect being tested (e.g., the relationship of the *random* effect to the *focal* effect in the basic design of the study). Thus, where s^2 stands for the observed, as opposed to hypothetical variances (σ^2), if random effect (r) is *nested* within focal effect (f):

$$F = s_f^2 / s_{r(f)}^2 \quad (2)$$

For example, in the case of the traditional between-subjects ANOVA design, the variance *between* experimental groups is divided by the variance among individuals nested *within* those groups. Individuals nested within groups represent the random effect in this model and hence the correct denominator for the *F*-ratio.

In contrast, if random effect (r) is *crossed* with focal effect (f):

$$F = s_f^2 / s_{r \times f}^2 \quad (3)$$

For example, in the traditional within-subjects ANOVA design, *treatments* are typically modeled as fixed effects, *subjects* as random effects, and treatments are fully crossed with subjects. The treatment effect is therefore tested against the *treatment * subjects* interaction, representing the variance in treatment effect across multiple subjects. Individual subjects crossed with groups represent the random effect in this model and hence the interaction term is the correct denominator for the *F*-ratio.

If one follows the logic closely, it is evident that these *F*-ratios are based upon the following *expected mean squares* (EMS) analysis of the hypothetical variance components underlying the *observed mean squares* (variances). To get directly to the conceptual heart of the issue, we momentarily dispense with the numerical weighting coefficients representing the relative numbers of experimental groups and numbers of subjects, which do little but add visual clutter and always balance out anyway in the equations. Thus, if r is *nested* within f:

$$s_f^2 = (\sigma_f^2 + \sigma_e^2) \quad (4)$$

and

$$s_{r(f)}^2 = \sigma_e^2 \quad (5)$$

Alternatively, if r is *crossed* with f:

$$s_f^2 = (\sigma_f^2 + \sigma_e^2) \quad (6)$$

and

$$s_{r \times f}^2 = \sigma_e^2 \quad (7)$$

These EMS analyses bring us full circle to the fundamental logic of the *F*-ratio described above.

As is no doubt well-known to the current readership of this special issue, the basic problem in much psycholinguistic research is that we often lack an “observed mean squares” term to use as the correct denominator in such an *F*-ratio. This is because we are often seeking to generalize to two hypothetical populations at the same time: that of *subjects* and that of *items*. The development of the quasi-*F*-ratio was an early attempt to construct the appropriate error terms for our treatment effects by judiciously using EMS analysis to algebraically produce the *F*-ratio denominators we need from those we are provided with by conventional ANOVA methods. The problems associated with the construction of quasi-*F*-ratios have been debated extensively elsewhere, so we will not review them here (Clark, 1973; Forster & Dickinson, 1976).

Instead, the present commentary is written to address the following question: If one is truly concerned about generalizing across samples of test items as well as samples of individual subjects, shouldn't one be using Generalizability Theory (GT), which was expressly design for such a function (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), rather than repeated measures ANOVA? Both are, at least implicitly, variance component models, which all the quasi-*F*-ratios are based on anyway. In contrast, GT explicitly permits the computation of separate generalizability coefficients across different *facets* (e.g., items and persons) of the design. It would seem that this approach would address the experimental questions more directly and provide easily interpretable coefficients of generalization, ranging between zero and one, which would doubtless be of more substantive value than a probability under the null hypothesis. Since GT allows one to generalize across different facets, one can still test F1 (i.e. generalize treatment across subjects) and F2 (i.e. generalize treatment across items), but can go beyond the normal null hypothesis significance testing (NHST) and parametrically estimate the magnitude of those effects.

A generalizability coefficient, which is a parameter estimate rather than a hypothesis test, is more informative than a statement that one has (or has not) rejected the null hypothesis. There are already many critiques of NHST in the literature (e.g., Cohen, 1990; Meehl, 1978, 1990), so we need not revisit that whole controversy in this commentary. The point is just that GT coefficients are pure parameter estimates and are thus not dependent on the implicit assumptions (e.g., random sampling) of NHST. They can be instead interpreted as estimates of the generalizability of the experimental effects across the sample of items used, and indirectly as representative of the theoret-

ical population of interest. This can be done without violating any of the assumptions of GT, and avoiding those of NHST entirely.

The logic of generalizability theory

The logic of the GT coefficient is ultimately based upon that of the “reliability” coefficient of Classical Test Theory (CTT). Because it provides a direct link to the logic of the *F*-ratio, we therefore review that first. The conceptual formula for a CTT reliability coefficient is as follows:

$$E^2 = \sigma_t^2 / (\sigma_t^2 + \sigma_e^2) \quad (8)$$

In this case σ_t^2 is referred to as the “true score” (rather than *treatment*) variance and σ_e^2 is still the “error” variance. Nevertheless, the use of a catch-all “error” term in CTT is deceptive. As we all know, there are various different CTT “reliability” coefficients possible (e.g., inter-item, inter-rater, test-retest) each seeking to generalize the true score across different dimensions (e.g., items, raters, testing occasions). This created a need for expanding our conceptualizations of reliability into the concept of generalizability.

Unlike Classical Test Theory, which compares a theoretically unitary “true score” to a homogeneous “error” term, Generalizability Theory (GT) recognizes *multiple* facets across which one may wish to generalize, called *random facets*, requiring alternative hypothesis tests and corresponding parameter estimates for the relative strengths of any generalizations across these multiple random facets. GT generalizability coefficients, based on estimated variance components, can be obtained by the following equations (Shavelson, Webb, & Rowley, 1989), where *f* is the *focal facet* and *r* is the *random facet*:

If *r* is *nested* within *f*:

$$E^2 = \sigma_f^2 / (\sigma_f^2 + \sigma_{r(f)}^2) \quad (9)$$

If *r* is *crossed* with *f*:

$$E^2 = \sigma_f^2 / (\sigma_f^2 + \sigma_{r \times f}^2) \quad (10)$$

Given these formulations, the relations of these coefficients to the *F*-ratios described above become more evident. For example, the classical repeated measures analysis of variance can be viewed as a special case of a random effects generalizability model for generalizing the effects of an experimental treatment on each subject across a random sample of subjects. In the traditional within-subjects design, *treatment* is typically modeled as a fixed effect and *subjects* as a random effect. The treatment effect is therefore tested against the *treatment * subjects* interaction, representing the variance in treatment effect across multiple subjects. The generalizability of the treatment effect across subjects would thus represent the proportion of treatment effect that was *invariant* between subjects, although this fascinating parameter estimate is not normally computed by traditional users of repeated measures analysis of variance. More mathematical treatments of this subject are presented in the ANOVA textbooks, involving the “intrusions” of the various “expected mean squares” generated by the crossing of random facets

with fixed. In contrast, GT theorists eschew tests of significance in favor of the estimation of variance component parameters. This is an eminently reasonable position because failing to reject the null hypothesis more often betokens a lack of sufficient statistical power than the absence of an effect in the real world (c.f., Cohen, 1990; Meehl, 1978, 1990).

An example of how these methods can be applied to categories of words is provided by Figueredo, Cox, and Rhine (1995). Although this was not a psycholinguistic study, it provided an analysis of a set of trait descriptors as applied to subjective personality assessments in nonhuman animals (e.g., monkeys and birds). The list of descriptive adjectives was divided into mutually exclusive categories based on previous exploratory factor analyses that indicated that they measured three distinct latent constructs (Stevenson-Hinde, Stillwell-Barnes, & Zunz, 1980; Stevenson-Hinde & Zunz, 1978). Using GT analyses, we assessed: (1) the convergent validities of these adjectives within the categories, (2) their reliabilities across multiple independent human raters, and (3) their temporal stability across individual (nonhuman animal) subjects across several years. This was all accomplished within a single GT model.

This multidimensional way of thinking is analogous to how similar ecological changes (e.g., in the dominant form of vegetation or *biomes*) can be observed when traveling higher in latitude (e.g., North) and when traveling higher in altitude (e.g., up a mountain). But these effects occur at different rates as a function of distance. For example, here in Southern Arizona, if you hike one mile up one of our local mountains (e.g., Mount Lemmon), you will eventually encounter coniferous forest biomes and maybe even snow at certain times of year. It would take you more like a thousand miles of hiking North, starting from the hot and dry Sonoran Desert, to reach a similar ecology if you remain within low altitudes. Thus, the relative magnitudes of generalizability coefficients across two different dimensions are of critical practical importance.

To demonstrate how this form of reasoning can be directly applied to problems in psycholinguistic research, we will temporarily abandon the ethereal realm of theory and present a concrete illustrative example of a real psycholinguistic experiment reanalyzed by means of GT. We will then return to theory and draw some general conclusions for future research.

An illustrative example

Lexical Inhibition theory explains that when an individual reads a word, they are searching for the correct word from a mental representation of a list of possible candidates; when they choose the most appropriate word, they are therefore simultaneously inhibiting other word choices. Researchers have tested Lexical Inhibition many times. The first type of experiment typically implemented uses a word very similar to the target as the prime (i.e., *able* as a prime for *AXLE*). This type of experiment usually (although not always) shows inhibition in lexical decision times. The second type of experiment uses a non-word

which is again very similar to the target as the prime (i.e., *aille* as a prime for *AXLE*). This type of experiment has usually, but again not always, shown facilitation in the lexical decision times. The inconsistencies in these results are demonstrated not only in the English language experiments, but across other languages as well. It is hypothesized that perhaps either having, or not having, a discrepancy in the frequency of the prime and target words is causing the inconsistencies. Davis and Lupker (2006) tested the effects of frequency discrepancy on lexical decision times for word primes, non-word primes, and unrelated control words.

The illustrative analyses in this commentary are based on Forster's as-yet unpublished (2007) replication of Davis and Lupker's (2006) study testing the effects of frequency discrepancies. Sixty target words (ITEM) were paired with a masked prime (60 ms) that was either a word (W) one letter different than the target, a non-word (NW) one letter different than the target, or a completely unrelated control (CONT) word. Half of the target words were higher (H) in frequency than their primes, and half were lower (L) in frequency. FREQUENCY does not refer to relative frequency, but instead could be described as frequency with a confound. That is, low-frequency targets were always primed with words higher in frequency, and high-frequency targets were always primed with words that were lower in frequency. The 60 target words were thus divided into the six experimental conditions, representing combinations of prime words and frequency discrepancies, resulting in 10 words per condition: H-W, H-NW, H-CONT, L-W, L-NW, and L-CONT. The order of these six conditions was systematically altered in their placement on three lists (LIST = A, B, C) to control for differences in the target frequency so that the targets were observed equally in each

prime condition across lists. See Table 1 for mean lexical decision times across PRIME and FREQUENCY.

Using Ordinary Least Squares (OLS) estimation, an analysis of variance can be performed which partitions the observed sums of squares, eta-squareds (proportions of the total variance accounted for by each source), and mean squares (variances) into all of its measured sources. This partitioning is shown in Table 2.

The traditional ANOVA approach is to construct the appropriate *F*-ratios and/or quasi-*F*-ratios from these observed mean squares. Based on the ANOVA, we see that the most amount of variance predicted is attributable to SUBJECT and SUBJECT * ITEM. However, we now depart from the conventional techniques.

The estimation of variance components is complex, but fortunately now can be performed through statistical programs such as SAS. As such, we will not discuss the estimation of variance components here, but will instead refer the reader to standard statistical texts such as Winer, Brown, & Michels (1991) or Bryk and Raudenbush (1992) for more detailed descriptions. Using the variance component estimation procedure in SAS (PROC VARCOMP), we can obtain a full Expected Mean Squares (EMS) analysis. For this illustrative analysis, we assumed the most complex ("worst case") scenario that all study facets were random effects. Although this might not have been the intention of the original authors, it will nonetheless permit us to show the flexibility of GT methods. This complete EMS analysis is shown in Table 3.

EMS analysis uses the fundamental principles behind ANOVA to construct equations representing all the hypothetical variance components, the so-called "expected mean squares", which theoretically contribute to the observed mean squares (variances). Because this breakdown is implicit in the experimental design, as are the weighting coefficients based on the numbers of treatment groups and the numbers of individual subjects in each condition, this part of the analysis can be performed automatically by the statistical software. Each of these "expected mean squares" equations is therefore set to be equivalent to the corresponding "observed mean squares" for each source of variance in the ANOVA model.

Given that the *observed* mean squares are known quantities, this system of simultaneous linear equations can

Table 1
Mean lexical decision times (ms) for prime by target frequency

Target frequency	Word	Non-word	Control
Low	579	577	591
	(hurt-HURL)	(nurl-HURL)	(eggs-HURL)
High	533	518	531
	(blur-BLUE)	(blae-BLUE)	(gasp-BLUE)

Table 2
Ordinary least squares (OLS) ANOVA table for illustrative study

Sources of variance	Degrees of freedom	Sums of squares	Eta-squareds (R-squareds)	Observed mean squares
LIST	2	690618	0.012	345309
SUBJECT	27	20699432	0.359	766646
FREQUENCY	1	1673578	0.029	1673578
LIST * FREQUENCY	2	10884	0.000	5442
SUBJECT * FREQUENCY	27	979600	0.017	36281
PRIME	2	5941	0.000	2970
LIST * PRIME	4	136870	0.002	34218
SUBJECT * PRIME	54	1635019	0.028	30278
FREQUENCY * PRIME	2	238083	0.004	119042
LIST * FREQUENCY * PRIME	4	77479	0.001	19370
SUBJECT * FREQUENCY * PRIME	54	1118367	0.019	20711
ITEM	54	3815362	0.066	70655
SUBJECT * ITEM	1293	26571112	0.461	20550
TOTAL	1526	57652346	1.000	

Table 3

Ordinary least squares (OLS) expected mean squares (EMS) analysis for illustrative study

Source	Expected mean squares	
LIST	σ^2 (SUBJECT * ITEM)	
	+0.0955 σ^2 (ITEM)	
	+8.7461 σ^2 (SUBJECT * FREQUENCY * PRIME)	
	+86.192 σ^2 (LIST * FREQUENCY * PRIME)	
	+0.1031 σ^2 (FREQUENCY * PRIME)	
	+17.134 σ^2 (SUBJECT * PRIME)	
	+170.09 σ^2 (LIST * PRIME)	
	+0.1527 σ^2 (PRIME)	
	+25.952 σ^2 (SUBJECT * FREQUENCY)	
	+257.77 σ^2 (LIST * FREQUENCY)	
	+0.0558 σ^2 (FREQUENCY)	
	+0.0558 σ^2 (FREQUENCY)	
	+51.091 σ^2 (SUBJECT)	
	+508.94 σ^2 (LIST)	
SUBJECT	σ^2 (SUBJECT * ITEM)	
	+0.1149 σ^2 (ITEM)	
	+8.7176 σ^2 (SUBJECT * FREQUENCY * PRIME)	
	+0.1116 σ^2 (LIST * FREQUENCY * PRIME)	
	+0.1116 σ^2 (FREQUENCY * PRIME)	
	+17.065 σ^2 (SUBJECT * PRIME)	
	+0.0673 σ^2 (LIST * PRIME)	
	+0.0673 σ^2 (PRIME)	
	+25.865 σ^2 (SUBJECT * FREQUENCY)	
	+0.1038 σ^2 (LIST * FREQUENCY)	
	+0.1038 σ^2 (FREQUENCY)	
	+50.878 σ^2 (SUBJECT)	
	FREQUENCY	σ^2 (SUBJECT * ITEM)
		+26.187 σ^2 (ITEM)
+83.763 σ^2 (LIST * FREQUENCY * PRIME)		
+251.01 σ^2 (FREQUENCY * PRIME)		
+0.132 σ^2 (LIST * PRIME)		
+0.338 σ^2 (PRIME)		
+250.34 σ^2 (LIST * FREQUENCY)		
+750.79 σ^2 (FREQUENCY)		
+8.5033 σ^2 (SUBJECT * FREQUENCY * PRIME)		
+0.1032 σ^2 (SUBJECT * PRIME)		
+25.174 σ^2 (SUBJECT * FREQUENCY)		
LIST * FREQUENCY		σ^2 (SUBJECT * ITEM)
		+0.1088 σ^2 (ITEM)
		+8.5003 σ^2 (SUBJECT * FREQUENCY * PRIME)
	+83.727 σ^2 (LIST * FREQUENCY * PRIME)	
	+0.1058 σ^2 (FREQUENCY * PRIME)	
	+0.104 σ^2 (SUBJECT * PRIME)	
	+0.1328 σ^2 (LIST * PRIME)	
	+0.0298 σ^2 (PRIME)	
	+25.163 σ^2 (SUBJECT * FREQUENCY)	
	+250.22 σ^2 (LIST * FREQUENCY)	
	SUBJECT * FREQUENCY	σ^2 (SUBJECT * ITEM)
		+0.1331 σ^2 (ITEM)
		+8.4508 σ^2 (SUBJECT * FREQUENCY * PRIME)
		+0.0911 σ^2 (LIST * FREQUENCY * PRIME)
+0.0911 σ^2 (FREQUENCY * PRIME)		
+0.1048 σ^2 (SUBJECT * PRIME)		
+0.1016 σ^2 (LIST * PRIME)		
+0.1016 σ^2 (PRIME)		
+25.011 σ^2 (SUBJECT * FREQUENCY)		
PRIME		σ^2 (SUBJECT * ITEM)
		+26.582 σ^2 (ITEM)
		+8.6571 σ^2 (SUBJECT * FREQUENCY * PRIME)
		+85.562 σ^2 (LIST * FREQUENCY * PRIME)
		+256.5 σ^2 (FREQUENCY * PRIME)
	+16.973 σ^2 (SUBJECT * PRIME)	
	+168.71 σ^2 (LIST * PRIME)	
	+505.85 σ^2 (PRIME)	

Table 3 (continued)

Source	Expected mean squares		
LIST * PRIME	σ^2 (SUBJECT * ITEM)		
	+0.0941 σ^2 (ITEM)		
	+8.6511 σ^2		
	(SUBJECT * FREQUENCY * PRIME)		
	+85.494 σ^2		
	(LIST * FREQUENCY * PRIME)		
	+0.0238 σ^2 (FREQUENCY * PRIME)		
	+16.958 σ^2 (SUBJECT * PRIME)		
	+168.57 σ^2 (LIST * PRIME)		
	SUBJECT * PRIME	σ^2 (SUBJECT * ITEM)	
		+0.1196 σ^2 (ITEM)	
		+8.6073 σ^2	
		(SUBJECT * FREQUENCY * PRIME)	
		+0.0669 σ^2	
(LIST * FREQUENCY * PRIME)			
+0.0669 σ^2 (FREQUENCY * PRIME)			
+16.851 σ^2 (SUBJECT * PRIME)			
FREQUENCY * PRIME		σ^2 (SUBJECT * ITEM)	
		+25.962 σ^2 (ITEM)	
		+8.332 σ^2	
		(SUBJECT * FREQUENCY * PRIME)	
		+82.569 σ^2	
		(LIST * FREQUENCY * PRIME)	
	+247.5 σ^2 (FREQUENCY * PRIME)		
	LIST * FREQUENCY * PRIME	σ^2 (SUBJECT * ITEM)	
		+0.1099 σ^2 (ITEM)	
		+8.3219 σ^2	
		(SUBJECT * FREQUENCY * PRIME)	
		+82.464 σ^2	
		(LIST * FREQUENCY * PRIME)	
		SUBJECT * FREQUENCY * PRIME	σ^2 (SUBJECT * ITEM)
+0.1414 σ^2 (ITEM)			
+8.2415 σ^2			
(SUBJECT * FREQUENCY * PRIME)			
ITEM			σ^2 (SUBJECT * ITEM)
			+24.944 σ^2 (ITEM)
SUBJECT * ITEM			σ^2 (SUBJECT * ITEM)

therefore be solved to obtain estimates of the hypothetical *expected* means squares comprising them. The EMS equations are essentially weighted sums of hypothetical variance components of unknown magnitude that theory tells us must be equal to their corresponding observed variances. Solving such a system of simultaneous linear equations is essentially identical to the way that chemists originally inferred the atomic weights of individual elements by knowing: (1) the molecular weights of their compounds and (2) the fixed combining proportions in which the elements were present in those compounds. For example, knowing that the formula for glucose is C₆H₁₂O₆, that of water is H₂O, and that of carbon dioxide is CO₂ permits one to infer the atomic weights of Carbon (C), Hydrogen (H), and Oxygen (O) from the weight of glucose, water and carbon dioxide without ever having purified or isolated these component elements. One can balance the equations for either respiration or photosynthesis and obtain the atomic weights of all component elements involved. By analogy, the *observed* mean squares are like the *molecules* of our compounds and the *expected* mean squares are like the *atoms* of our elements. As with many elements, *expected* mean squares are not observed in isola-

tion, but only in various combinations. Because statistical theory tells us in exactly what numerical proportions the hypothetical variance components constitute our observed variances, we can solve the simultaneous system of equations to obtain our best estimates for them.

There are two estimation methods which can be applied towards estimating the hypothetical variance components. The first method, referred to as Ordinary Least Squares (OLS) estimation, works by finding a model that best fits the data by minimizing the unweighted sum of squared errors (Field, 2005). We therefore applied OLS estimation, as we did to obtain the original ANOVA results, using the METHOD = TYPE 1 option within PROC VARCOMP for hierarchical OLS estimation, and obtained the results displayed in Table 4.

This result, however, is unsatisfactory from a theoretical standpoint because it produces several negative estimates for certain variance components, which are not theoretically permissible. This should serve as a caution that the conventional algebraic computations of quasi-*F*-ratios, although perhaps soundly based on statistical theory, might be incorrectly estimated when based on variances obtained from the OLS estimation normally used in ANOVA, which might be producing results that are implicitly out of permissible bounds according to statistical theory.

To solve this problem, we switch to Restricted Maximum Likelihood (REML) estimation, using the METHOD = REML option within PROC VARCOMP, which constrains estimates for variance components to be no lower than zero. REML, a type of Maximum Likelihood (ML) estimation, is more applicable with small sample sizes (a typical cause of negative variance components) (Zhang, Johnston, & Kilic, 2007). REML and other ML methods differ from OLS because they base their estimates on probability theory. ML estimation assumes that the data collected is a smaller sample of a larger population and, using probability theory, iteratively estimates the likelihood that what happened in the sample will occur in the larger population (Harell, 2001). REML differs slightly from full ML in its statistical algorithms and capabilities. ML estimates random components as well as fixed Level II coefficients by maximizing the joint likelihood between them, whereas REML estimates random components by averaging them over all possible values of fixed effects (Raudenbush, 2001).

REML is recommended over ML for estimating hypothetical variance components because it estimates the variances and covariances based on intervals instead of using fixed estimates from the fixed effects. Thus, REML is considered more realistic because it is capable of adjusting for uncertainty around the fixed effects (Tabachnick & Fidell, 2007). The REML results for these data are displayed in Table 5. These estimates are now no longer out of permissible bounds and can be interpreted further.

Another procedure that can be used for this kind of analysis is Hierarchical Linear Modeling (HLM), which can be accomplished within SAS using the PROC MIXED procedure, which can handle any combination of random and fixed effects within the same model. Although it is beyond the scope of this commentary to detail that procedure here, we will mention that HLMs are variance component models based upon the same fundamental theory as GT analyses and that variance component estimates can be obtained from HLM output. The two approaches are therefore not mutually exclusive, but may be viewed as providing complementary information. A detailed description of how all of the relevant variance components can be estimated within HLM is to be found in Bryk and Raudenbush (1992).

We see that several sources of variance in our original design, such as LIST and PRIME, indeed have estimated variance components of zero. Others have non-zero variance components of varying magnitudes. Recall that the entire rationale for constructing *F*-ratios or quasi-*F*-ratios was to identify which variance components were (or were not) equal to zero. Variance component estimation tells us that directly. Although Generalizability theorists avoid the jargon of NHST, these effects could be considered “non-significant” by the conventional logic of ANOVA. However, GT analysis can go beyond NHST and tell us more.

For example, applying the general formulas for GT coefficients provided above, we may construct a variety of interesting GT coefficients for the non-zero effects. Some interesting coefficients that can be computed are shown in Table 6. As mentioned earlier, these GT coefficients represent the proportion of the facet of interest that was invariant across the error term. For example, the generalizability of treatment across subjects represents the propor-

Table 4

Ordinary least squares (OLS) variance component (VARCOMP) estimation for illustrative study

Variance component	Estimates
σ^2 (LIST)	-782
σ^2 (SUBJECT)	14151
σ^2 (FREQUENCY)	2088
σ^2 (LIST * FREQUENCY)	-118
σ^2 (SUBJECT * FREQUENCY)	621
σ^2 (PRIME)	-265
σ^2 (LIST * PRIME)	31
σ^2 (SUBJECT * PRIME)	570
σ^2 (FREQUENCY * PRIME)	193
σ^2 (LIST * FREQUENCY * PRIME)	-15
σ^2 (SUBJECT * FREQUENCY * PRIME)	-15
σ^2 (ITEM)	2009
σ^2 (SUBJECT * ITEM)	20550

Table 5

Restricted maximum likelihood (REML) variance component (VARCOMP) estimation for illustrative study

Variance component	Estimates
σ^2 (LIST)	0
σ^2 (SUBJECT)	13568
σ^2 (FREQUENCY)	2470
σ^2 (LIST * FREQUENCY)	0
σ^2 (SUBJECT * FREQUENCY)	470
σ^2 (PRIME)	0
σ^2 (LIST * PRIME)	0
σ^2 (SUBJECT * PRIME)	533
σ^2 (FREQUENCY * PRIME)	0
σ^2 (LIST * FREQUENCY * PRIME)	0
σ^2 (SUBJECT * FREQUENCY * PRIME)	0
σ^2 (ITEM)	2227
σ^2 (SUBJECT * ITEM)	20578

Table 6

Restricted maximum likelihood (REML) generalizability coefficients for illustrative study

Focal facet	Random "error" facet	G-coefficient
FREQUENCY	Across subjects	SUBJECT * FREQUENCY .840
	Across primes	FREQUENCY * PRIME 1.000
	Across items	ITEM .526
SUBJECT	Across frequencies	SUBJECT * FREQUENCY .967
	Across primes	SUBJECT * PRIME .962
	Across items	ITEM .397
ITEM	Across subjects	SUBJECT * ITEM .098

tion of treatment effect that was invariant between subjects.

The first three coefficients shown indicate the generalizability of FREQUENCY across three alternative random facets: (1) across individual subjects, (2) across the different types of primes, and (3) across the different items sampled. The first coefficient of .84 indicates that the effect of FREQUENCY was quite generalizable across different subjects, meaning they all responded quite similarly to words of equivalent frequency. The second coefficient of 1.00 indicates that the effect of FREQUENCY generalized perfectly across the different types of primes, because the variance component estimated for the FREQUENCY * PRIME interaction was zero. The third coefficient of .53 indicates that the effect of FREQUENCY was somewhat generalizable across different items, but not as much as one might hope. This cautions that the choice of particular test words used, as psycholinguists have long suspected, is an important one. Furthermore, high variability among the effects of particular words indicates the need for an adequate sampling of items to estimate the central tendencies for the experimental conditions.

The second three coefficients shown indicate the generalizability of the SUBJECT effect across three alternative random facets: (1) across the different word frequencies, (2) across the different types of primes, and (3) across the different items sampled. These effects are not typically tested by psycholinguists because they are not manipulated experimental conditions referring to the properties of language but instead are observed systematic individual differences in reaction time to the presented stimuli. The first coefficient of .97 indicates that the effect of SUBJECT was quite generalizable across different word frequencies, meaning that systematic differences among individual subjects were surprisingly similar across words of different frequency. The second coefficient of .96 indicates that the effect of SUBJECT also generalized quite well across different types of primes. The third coefficient of .40, however, indicates that the effect of SUBJECT was not very generalizable across different items, indicating that individual subjects had somewhat idiosyncratic responses to different test words of equivalent frequency. These results emphasize the importance of systematic individual differences, which are something typically not addressed within the field of psycholinguistics, but nonetheless constitute the primary subject matter of other fields such as intelligence research. Indeed, reaction time is often used as a chronometric indicator of general intelligence.

The final coefficient shown indicates the generalizability of the ITEM effect across the random facet representing different individual subjects. The coefficient of just less than .10 indicates that this generalizability is extremely poor, again indicating a pattern of idiosyncratic responding to test words by individual subjects.

This analysis also showed no effect of PRIME, or FREQUENCY * PRIME meaning there was no difference found in response times between the different priming conditions, or in the interaction of the different priming conditions with high or low frequency. These results are contrary to what was originally predicted by Davis and Lupker (2006), and by Forster (2007). PRIME, however, tested all three priming conditions (i.e. word, non-word, and control) and did not initially make comparisons between specific prime conditions. To ascertain whether there may have been an effect between specific priming conditions when considered separately, we also ran a separate GT analysis which restricted PRIME to just *non-word* and *control*, temporarily removing the *word* prime condition from the analysis. In addition, we ran a separate GT analysis which restricted PRIME to just *word* and *control*, this time temporarily removing the *non-word* prime condition from the analysis. We still found no variance attributable to PRIME in either of these restricted pairwise comparisons. We believe that the fact that GT found no effect of PRIME when comparing non-word with control and that ANOVA did speak, not to the lack of sensitivity of GT, but to potential of oversensitivity of quasi-*F*-ratios.

GT analyses which compare only two conditions of a specific facet allow one to test specific hypotheses of how that particular condition may generalize over other facets of the study such as items or subjects. For example, if one wanted to test whether the lack of priming is due to large individual differences, one would generalize PRIME over PRIME * SUBJECTS. PRIME * SUBJECTS would be the interindividual variance between subjects on priming conditions. If PRIME * SUBJECTS had a large variance component compared with PRIME, then the generalizability of PRIME would be low. If PRIME * SUBJECTS had a small variance component compared with PRIME, then the generalizability of PRIME would be high. One could assume that the effects of priming might be larger if there was not so much variability between participants on priming conditions. This same example could be applied to items as well so if PRIME * ITEMS had a large variance component, then the generalizability of PRIME would be low.

What these extra coefficients demonstrate is the great flexibility of GT analyses to provide more information than is usually available with more conventional techniques. For example, separate item analyses are often necessary using conventional methods, whereas GT methods allows the independent exploration of each separate facet of the study using generalizability coefficients computed along different dimensions. These coefficients allow one to compare the relative magnitudes of the effects within a common metric and not just whether each one is "statistically significant", which just means greater than zero (or "better than nothing!").

Discussion

Sir Ronald Fisher, in a seldom-heeded caveat, cautioned that NHST should serve as the beginning of a scientific exploration of a phenomenon and not as its end. As Meehl (1978, 1990) has pointed out, however, contemporary publication practices have rendered the “tabular asterisks” (indicating .05 significance, which was a relatively arbitrary decision rule also proposed by Fisher with little solid rationale behind it) the “coin of the realm” in peer review. To correct this bias, others such as Cohen (1990) have proposed that we should instead emphasize effect size estimation over NHST to indicate the relative importance of our results, and more routinely use power analysis to put the results our significance tests in better perspective, since they are so critically sensitive to the statistical power actually involved (for example, Meehl had shown that the .05 statistical criterion was actually met about 50% of the time using a sufficiently large sample of real data!).

Generalizability Theory (Cronbach et al., 1972) provides us a way out of this quandary. The major claim in this commentary is that it is especially useful in psycholinguistics research because it affords us the opportunity to assess generalizability across multiple dimensions (random effects) within the same model. This can be accomplished without having to construct elaborate quasi-*F*-ratios. This commentary does not in any way challenge the mathematical *validity* of these quasi-*F*-ratios. What we question is the practical *utility* of creating a single catch-all “error” term as the denominator. Such a ratio implicitly estimates the overall generalizability of the focal effect across multiple dimensions at once, and is therefore not very informative. We suggest that it is better to be more specific and determine the separate and perhaps distinct degrees of generalizability of a given effect across any number of independent dimensions (e.g., subjects, words, etc.).

In this commentary, we have used an illustrative empirical example from contemporary psycholinguistic research to demonstrate how the principles of GT can be productively applied to circumvent the difficulties posed by the multiple-random-effects designs that are so prevalent in this field. We have not directly addressed the merits and

limitations of quasi-*F*-ratios but instead provided a way to circumvent them. We hope that this provides a useful alternative approach for researchers in the field.

References

- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park: Sage.
- Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12, 335–359.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 1304–1312.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York, NY: Wiley.
- Davis, C. J., & Lupker, S. J. (2006). Masked inhibitory priming in English: Evidence for lexical inhibition. *Journal of Experimental Psychology: Human Perception and Performance*, 32(3), 668–687.
- Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London: SAGE Publications.
- Figueredo, A. J., Cox, R. L., & Rhine, R. J. (1995). A generalizability analysis of subjective personality assessments in the Stumptail macaque and the Zebra finch. *Multivariate Behavioral Research*, 30(2), 67–197.
- Forster, K. I., & Dickinson, R. G. (1976). More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F1, F2, F, and minF. *Journal of Verbal Learning and Verbal Behavior*, 15, 135–142.
- Harell, F. E. (2001). *Regression modeling strategies with applications to linear models, logistic regression, and survival analysis*. New York: Springer-Verlag.
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806–834.
- Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66, 195–244.
- Raudenbush, S. W. (2001). Toward a coherent framework for comparing trajectories of individual change. In L. M. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 33–64). Washington, DC: American Psychological Association.
- Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, 44, 922–932.
- Stevenson-Hinde, J., Stillwell-Barnes, R., & Zunz, M. (1980). Subjective assessment of rhesus monkeys over four successive years. *Primates*, 21(1), 66–82.
- Stevenson-Hinde, J., & Zunz, M. (1978). Subjective assessment of individual rhesus monkeys. *Primates*, 19, 473–482.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson.
- Winer, B. J., Brown, R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York, NY: McGraw-Hill.
- Zhang, B., Johnston, L., & Kilic, G. B. (2007). Assessing the reliability of self and peer rating in student group work. *Assessment and Evaluation in Higher Education*, 1–12.