Shiken: JALT Testing & Evaluation SIG Newsletter. 12 (2) April 2008 (p. 38 - 43)

Statistics Corner
Questions and answers about language testing statistics:

Effect size and eta squared

James Dean Brown (University of Hawai'i at Manoa)

Question: In Chapter 6 of the 2008 book on heritage language learning that you co-edited with Kimi-Kondo Brown, a study comparing how three different groups of informants use intersentential referencing is outlined. On page 147 of that book, a MANOVA with a partial eta^2 of .29 is outlined. There are several questions about this statistic. What does a "partial eta" measure? Are there other forms of eta that readers should know about? And how should one interpret a partial eta^2 value of .29?

Answer: I will answer your question about partial eta^2 in two parts. I will start by defining and explaining *eta*^{2}. Then I will circle back and do the same for *partial eta*^{2}.

Eta^2

Eta^2 can be defined as the proportion of variance associated with or accounted for by each of the main effects, interactions, and error in an ANOVA study (see Tabachnick & Fidell, 2001, pp. 54-55, and Thompson, 2006, pp. 317-319). Formulaically, eta^2, or η^2, is defined as follows:

$$\eta^2 = \frac{SS_{effect}}{SS_{total}}$$

Where:

SS_{effect} = the sums of squares for whatever effect is of interest
SS_{total} = the total sums of squares for all effects, interactions, and errors in the ANOVA

Eta^2 is most often reported for straightforward ANOVA designs that (a) are balanced (i.e., have equal cell sizes) and (b) have independent cells (i.e., different people appear in each cell). For example, in Brown (2007), I used an example ANOVA to demonstrate how to calculate power with SPSS. That was a 2 x 2 two-way ANOVA with *anxiety* and *tension* as the independent variables and *trial 3* as the dependent variable (using the *Anxiety 2.sav* example file that comes with recent versions of the *SPSS* software). There were three people in each cell and the cells were independent.

Notice in Table 1 that the *p* values (0.90, 0.55, & 0.10) indicate that there were no significant effects (i.e., no *p* values below .05) for Anxiety, Tension, or their interaction. Note also that there was not sufficient power to detect such effects (i.e., the power statistics of 0.05, 0.09, & 0.37 were not above .80 in any case). All of this led me to conclude that "the study lacked sufficient power to detect any significant effects even if they exist in reality", which is reasonable given the very small sample size of 12.

Table 1 *Results of the Analysis Shown in Figure 3 of the Anxiety 2.sav used with SPSS*

Source	*SS*	*df*	*MS*	*F*	*p*	eta^2	Power
Anxiety	0.08	1	0.08	0.02	0.90	0.0012	0.05
Tension	2.08	1	2.08	0.38	0.55	0.0324	0.09
Anxiety x Tension	18.75	1	18.75	3.46	0.10	0.2919	0.37
Error	43.33	8	5.42			0.6745	
Total	64.24	12					

Table 2 *Descriptive Statistics for the Anxiety 2.sav Example Used with SPSS* *

Anxiety	Tension	*M*	*SD*	*N*
1	1	8.67	3.06	3
	2	7.00	2.65	3
2	1	6.00	2.00	3
	2	9.33	1.16	3

*Dependent Variable: Trial 3

Nonetheless, even a cursory look at the means shown in Table 2 indicates that fairly large differences exist between means and something noteworthy is going on, so a better designed replication study with a larger sample size might be justified. Eta^2 can help in interpreting the results by indicating the relative degree to which the variance that was found in the ANOVA was associated with each of the main effects (Anxiety and Tension) and their interaction.

Eta^2 values are easy to calculate. Simply add up all the sums of squares (*SS*), the total of which is 64.24 in the example; then, divide the *SS* for each of the main effects, the interaction, and the error term by that total. The results will be as follows:

$$\eta^2_{Anxiety} = \frac{SS_{Anxiety}}{SS_{Total}} = \frac{0.08}{64.24} = 0.00124533 \approx 0.0012$$

$$\eta^2_{Tension} = \frac{SS_{Tension}}{SS_{Total}} = \frac{2.08}{64.24} = 0.03237858 \approx 0.0324$$

$$\eta^2_{AxT} = \frac{SS_{AxT}}{SS_{Total}} = \frac{18.75}{64.24} = 0.291874221 \approx 0.2919$$

$$\eta^2_{Error} = \frac{SS_{Error}}{SS_{Total}} = \frac{43.33}{64.24} = 0.674501867 \approx 0.6745$$

Interpretation of these values is easiest if the decimal point is moved two places to the right in each case, the result of which can be interpreted as percentages of variance associated with each of the main effects, the interaction, and error.

Starting with Anxiety, the value of 0.0012 indicates that a mere 0.12% of the variance is accounted for by Anxiety, whereas Tension accounts for 3.24%, the Anxiety x Tension (A x T) interaction accounts for a much larger 29.19%, and a whopping 67.45% is accounted for by Error. Now let's consider the A x T interaction and Error separately in more detail.

The 29.19% accounted for by the A x T interaction should lead the researcher to understand that this interaction effect is much more important than either of the individual main effects for Anxiety or Tension, a fact that, even though there are no significant effects, may help in designing future studies and understanding why the present one did not detect significant differences. Such an important interaction effect should lead the researcher to want to plot out that relationship as shown in Figure 1, where we see that the Tension groups 1 (dotted line) and 2 (plain black line) do indeed have different means but in opposite relationships for Anxiety 1 and 2. That is, the Tension 1 group is higher than the Tension 2 group when Anxiety is 1, but the Tension 1 group is lower than the Tension 2 group when Anxiety is 2.

Thus there is a strong pattern but it is not consistent across Anxiety 1 and 2 conditions (if it were consistent, the lines would be parallel). Thus, even with a non-significant interaction (where $p = .10$), the eta^2 value of .2919 drew our attention to an important interaction effect that is revealing in itself, and which may help to understand why there were no significant main effects for Tension or Anxiety (i.e., because the interaction cancels out any such differences).

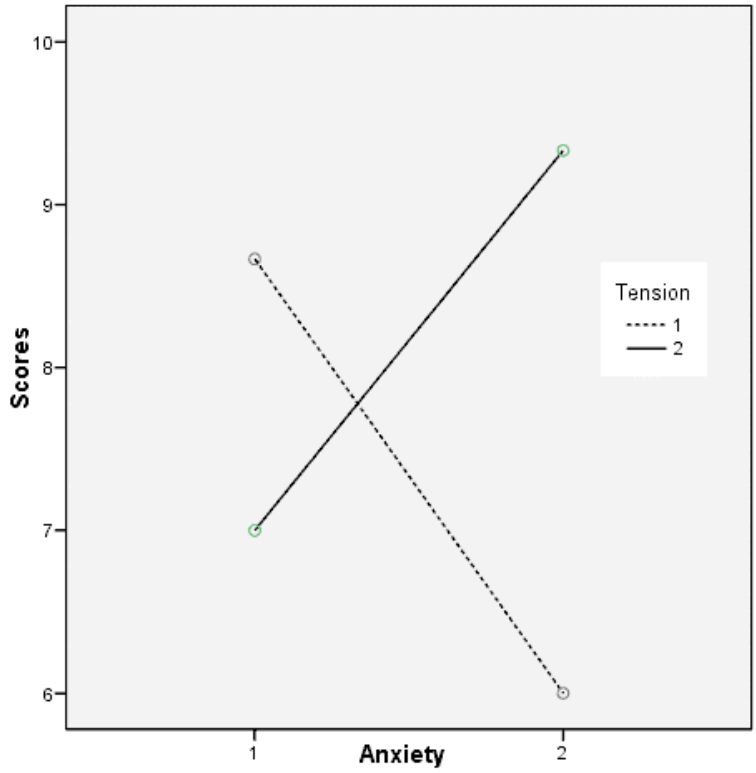

Figure 1. Interaction of Anxiety with Tension using the *Anxiety 2.sav* example

The whopping 67.45% accounted for by *Error* in the Table 1 indicates that more than two-thirds of the variance was not accounted for at all in this design. This error variance may be due to unreliable variance in the study due to poor design, other systematic variables that might be of interest (if they were operationalized and included in the study), and so forth. All in all, eta^2 values indicate not only that the interaction effect and error are causing almost 97% of the variance in the study (67.45 + 29.19 = 96.64), but also ways to redesign the study so it will be more powerful and meaningful.

"One problem with eta^2 is that the magnitude of eta^2 for each particular effect depends to some degree on the significance and number of other effects in the design . . ."

One problem with eta^2 is that the magnitude of eta^2 for each particular effect depends to some degree on the significance and number of other effects in the design (Tabachnick & Fidell, 2001, p. 54). One statistic that minimizes the effects of this issue is called *partial eta^2*.

Partial Eta^2

Partial eta^2 can be defined as the ratio of variance accounted for by an effect and that effect plus its associated error variance within an ANOVA study. Formulaically,

partial eta^2, or $\eta^2_{partial}$, is defined as follows:

$$\eta^2_{partial} = \frac{SS_{effect}}{SS_{effect} + SS_{error}}$$

Where:

SS_{effect} = the sums of squares for whatever effect is of interest

SS_{error} = the sums of squares for whatever error term is associated with that effect

In applied linguistics studies, partial eta^2 is most often reported for ANOVA designs that have non-independent cells (i.e., the same people appear in more than one cell). For example, in Brown, Hilgers, and Marsella (1991), students wrote compositions on two different types of topics (a narrative topic and an analytic topic) which were organized into ten prompt sets. The people who wrote on each of the ten prompt sets were different from each other (so this is also known as a *between subjects effect*). In contrast, every student wrote on each of the two topic types, so these were treated as repeated measures (also known as a *within subjects effect*). The cell sizes *within subjects* were exactly the same (which makes sense because they were the same people), whereas the cell sizes *between subjects* were different to small degrees. The original results of this 10 x 2 two-way repeated-measures ANOVA for prompt sets and topic types are shown in Table 3.

Table 3 *Two-Way Repeated-Measures ANOVA for 1989 Prompt Sets and Topic Types (As presented in Brown et al, 1991)*

Source	*SS*	*df*	*MS*	*F*	*p*
Between Subjects					
Prompt Set	158.372	9	17.597	9.703	0.00
Error	3068.553	1692	1.814		
Within Subjects					
Topic Type	0.344	1	0.344	0.194	0.66
Prompt Set by Topic Type	137.572	9	15.286	8.611	0.00
Error	3003.548	1692	1.775		

Table 4 *Two-Way Repeated-Measures ANOVA for 1989 Prompt Sets and Topic Types (Adapted from Brown et al, 1991 with Partial Eta^2 Added)*

Source	*SS*	*df*	*MS*	*F*	*p*	Partial eta^2
Between Subjects						
Prompt Set (PS)	158.372	9	17.597	9.703	0.00	0.0490
$Error_{BS}$	3068.553	1692	1.814			
Within Subjects						
Topic Type (TT)	0.344	1	0.344	0.194	0.66	0.0001
Prompt Set by Topic Type	137.572	9	15.286	8.611	0.00	0.0438
$Error_{WS}$	3003.548	1692	1.775			

From my present perspective (17 years later), the 1991Brown et al. study would have been strengthened by relabeling the effects and adding partial eta^2 values to the two-way repeated-measures ANOVA table as shown in Table 4. These partial eta^2

values are easy to calculate. Simply divide the *SS* for each effect by the *SS* of that effect plus the *SS* for the error associated with that effect. The results will be as follows:

$$\text{Partial } \eta^2_{PS} = \frac{SS_{PS}}{SS_{PS} + SS_{Error_{BS}}} = \frac{158.372}{158.372 + 3068.553} = 0.049078302 \approx 0.0490$$

$$\text{Partial } \eta^2_{TT} = \frac{SS_{TT}}{SS_{TT} + SS_{Error_{WS}}} = \frac{0.344}{0.344 + 3003.548} = 0.000114518 \approx 0.0001$$

$$\text{Partial } \eta^2_{PSxTT} = \frac{SS_{PSxTT}}{SS_{PSxTT} + SS_{Error_{WS}}} = \frac{137.572}{137.572 + 3003.548} = 0.043797116 \approx 0.0438$$

The interpretation of these partial eta^2 values is similar to what we did above for eta^2 in that we need to move the decimal point two places to the right in each case, and interpret the results as percentages of variance. However, this time the results indicate the percentage of variance in each of the effects (or interaction) and its associated error that is accounted for by that effect (or interaction). Starting with Prompt Sets, the value of 0.0490 indicates that 4.90% of the between subjects variance is accounted for by Prompt Sets, whereas Topic Types accounts for nearly none of the TT plus Error$_{BS}$ variance (0.01%), though the Prompt Sets by Topic Types interaction (PSxTT) accounts for a somewhat larger 4.38% of the PSxTT plus Error$_{BS}$ variance.

Conclusion

In direct answer to your question, Kondo-Brown and Fukuda (2008) correctly chose to use partial eta^2 because their design was a MANOVA, which by definition involves non-independent or repeated measures. When they reported that partial eta^2 was .29, that meant that the effect for group differences in their MANOVA accounted for 29% of the group-differences plus associated error variance as explained above. This percentage was sufficient to lead them to do univariate follow-up ANOVAs that helped them to further isolate exactly where the significant and interesting means differences were to be found.

In recent columns, I have covered a number of issues related to the ANOVA sorts of studies including: sampling and generalizability, sampling errors, sample size and power, and effect size and eta squared. All of these are ways to expand your thinking about ANOVA—ways that are often ignored in applied linguistics. They have long been important to understanding ANOVA results in psychology, education, and other fields, and we ignore them to our detriment. To paraphrase something one of my stats teachers said back in the late 1970s: Reporting the traditional ANOVA source table (with *SS*, *df*, *MS*, *F*, and *p*) and discussing the associated significance levels isn't the *end* of the study; it's just the *beginning* because we can learn much more by carefully plotting and considering the interaction effects and doing follow up analyses like planned or post-hoc comparisons, power and effect size analyses, and so forth. I hope I have delivered that message loud and clear.

"Reporting the traditional ANOVA source table (with SS, df, MS, F, and p) and discussing the associated significance levels isn't the end of the study; it's just the beginning . . ."

References

Brown, J. D. (2007). Statistics Corner. Questions and answers about language testing statistics: Sample size and power. *Shiken: JALT Testing & Evaluation SIG Newsletter, 11*(1), 31-35. Also retrieved from the World Wide Web at http://jalt.org/test/bro_25.htm

Brown, J. D., Hilgers, T., & Marsella, J. (1991). Essay prompts and topics: Minimizing the effect of differences. *Written Communication, 8*(4), 532-555.

Kondo-Brown, K. (2005). Differences in language skills: Heritage language learner subgroups and foreign language learners. *The Modern Language Journal, 89*(4), 563–581.

Kondo-Brown, K., & Brown, J. D. (Eds.) (2008). *Teaching Chinese, Japanese, and Korean heritage language students.* New York: Lawrence Erlbaum Associates.

Kondo-Brown, K., & Fukuda, C. (2008). A separate-track for advanced heritage language students?: Japanese intersentential referencing. In K. Kondo-Brown & J. D. Brown (Eds.), *Teaching Chinese, Japanese, and Korean heritage language students.* New York: Lawrence Erlbaum Associates.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics (5th ed.)*. Upper Saddle River, NJ: Pearson Allyn & Bacon.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York: Guilford.

HTML: http://jalt.org/test/bro_28.htm / PDF: http://jalt.org/test/PDF/Brown28.pdf