

Generality of effects in item analyses

Florian Schwarz
Department of Linguistics
University of Massachusetts at Amherst

In a recent discussion of the limits of F_2 's for psycholinguistic analyses, Forster (2006) suggests that we need a measure of generality of some sort, which tells us something about the extent to which an effect that we have found is carried by all or most items. This is typically of importance for psycholinguistic theorizing, as our hypotheses are usually about entire classes of items (or expressions), and not about some specific subset thereof. As Forster (2006) shows, F_2 's cannot answer this question, since one easily obtains significant p-values for an effect that is carried by a small number of outliers. As a remedy for this, Forster suggests a procedure of iteratively removing the item displaying the strongest effect and taking the proportion of items that can be removed while still obtaining a significant p-value as a measure of generality. In the following, I sketch an alternative measure of generality that yields comparable results but which possibly could be based on more refined statistical theory by statisticians (at least this is the hope here).

The central idea is the following: For each item, we can evaluate how the effect for this specific item relates to the mean effect found for the entire data set by comparing the actual value in the treatment condition with the value that we would predict if we assumed the data point to be missing and calculated it using standard data imputation techniques. If an effect is perfectly general, this prediction matches the actual value. If the effect is produced by a few outliers with very strong effects, the predicted and actual values will differ substantially. How much they will differ depends on how unevenly the effect is distributed across items. By forming the sums of squares of differences between predicted and actual values, we can calculate a particular kind of standard deviation for the mean of the treatment condition and determine a 95 per cent confidence interval in the usual way, which in turn allows inferences about the generality of the mean effect. If the actual mean effect is bigger than two of these standard deviations of the mean for the treatment condition, we can conclude the effect to be general.

Let me begin the illustration of this idea by introducing the formula used for data imputation (from Keppel & Wickens 2004: 396):

$$\text{Estimated } Y_{ij} = \frac{aA_{j-} + nS_{i-} - T-}{(a-1)(n-1)}$$

a: # of conditions, A_{j-} : sum of scores in condition j, n: number of subjects, S_{i-} : sum of scores for subject i, T-: sum of all scores (all sums of scores exclude the score Y_{ij} , of course, since it is missing)

Using a very simple example, this formula works as follows:

	Conditions:	A	B
Items	1	10	-
	2	10	20
	3	10	20
	4	10	20

$$\text{Estimated } Y_{1B} = \frac{2*60 + 4*10 - 100}{(2-1)(4-1)} = 20$$

Of course, for present purposes, we are assuming that actually none of the data points are missing. What we are using the imputation procedure for is to determine to what extent the effect we found is due to particular items. For concreteness, let's assume that the data point missing above is 50. The following table illustrates the steps of the procedure I'm suggesting:

Item	Condition A	Condition B	Imputed score for B	Difference between B and imputed B	Squared difference
1	10	50	20	30	900
2	10	20	40	20	400
3	10	20	40	20	400
4	10	20	40	20	400
Mean	10	27.5			

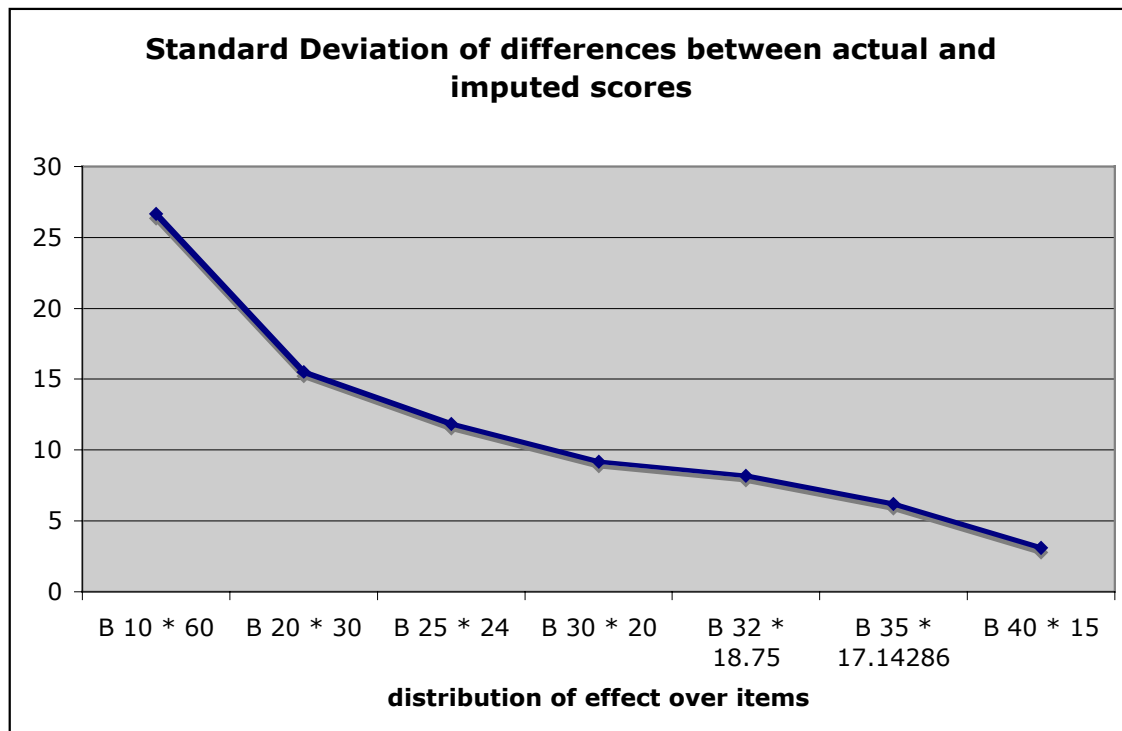
Sum of Squares: 2000
Mean Squares: 666.66
Standard Deviation: 25.8

In terms of the means, the effect is 17.5, but given the procedure I'm proposing, that difference is not a general one, due to the outlier of 50. Note that if we had assumed the data point missing above to be 20 rather than 50, the imputed scores would have been identical to the actual ones, which would have resulted in the differences and the sums of squares being 0. Further note that the result for the sum of squares obtained would have been identical if we had imputed scores for condition A rather than for B.

What does the result tell us, in conceptual terms. Speaking in terms of the standard deviation, it tells us something like the following: considering the possibility for each one of the data points that it would be absent, how would the overall effect have been affected by imputing that score? In this sense, it tells us to what extent the overall effect depended on any of the particular items chosen, so it should allow us to draw inferences about how general the overall effect is across the different items we have looked at. The standard

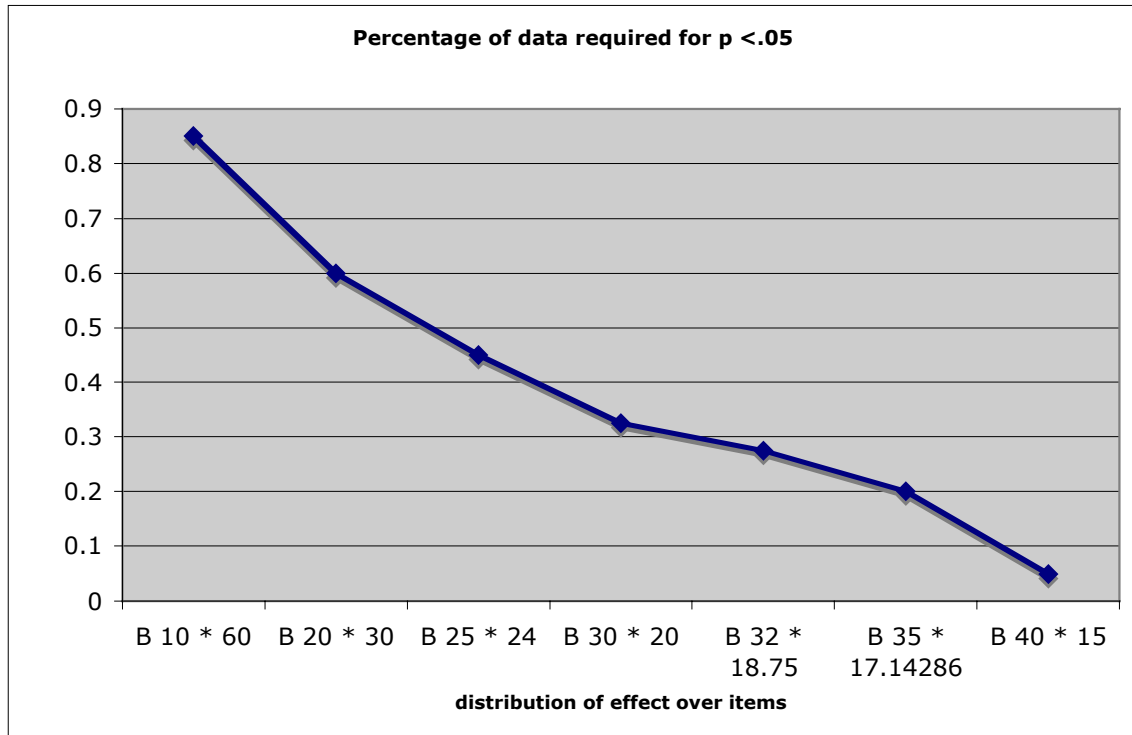
deviation of 25.8 tells us that the effect of 17.5 that we found depended rather strongly on our particular set of items. Speaking in terms of a 95 per cent confidence interval, it tells us that the general mean for the treatment condition lies within the range of +/- 51.6 (i.e. 2 SD) of the mean of 27.5. This interval includes the mean of condition A, hence we cannot make any conclusions about the effect we found being a general one.

Having looked at an extremely simple case for illustration of the basic idea, we can turn to discussing some more realistic scenarios. Following the method in Forster (2006), 40 data points with a mean of 583 and a standard deviation of 68 were randomly generated, and from this a second set of 40 data points was generated by adding a random component with a mean of 0 and a variance of 10 to the corresponding values in the first set. A total effect of 600 was then added to 10, 20, 25, 30, 32, 35, and 40 of the items (10 * 60, 20 * 30, etc.). As in Forster (2006), this led to highly significant differences in means even in the case where only 10 items carried the effect. However, a different picture emerges when we look at the standard deviations for the differences between actual and imputed scores. As the following graph illustrates, the SD decreases as the effect is distributed more evenly over the data.



Considering the actual effect in the data which was 14.44, the threshold of 2 SD is reached somewhere between 32 and 35 items that the effect is distributed over. That is, for an effect to be general according to this procedure, it has to be distributed at least across 80-87.5 per cent of the data (assuming perfectly even distribution of the effect across these items).

It turns out that this curve is remarkably similar to a curve obtained from applying Forster's procedure of calculating p-values after iteratively removing the data point with the largest effect in the remaining set. Mapping this for the same data by representing the percentage of data necessary for obtaining a p-value smaller than .5 yields the following graph:



The similarity between the two graphs is rather striking. Hopefully, more sophisticated statisticians can figure out how these two approaches are related mathematically. It seems to me like the procedure proposed here is less crude and more mathematically accurate (for example, it should be more accurate when dealing with smaller sets of items, where each removed item counts for a larger percentage of the entire set), but it remains to be seen whether the conceptual motivation for that procedure sketched here can be fully fleshed out by statisticians.