A Further Look at the "Language-as-Fixed-Effect Fallacy"

Abstract The proper analysis of experiments using language materials has been a source of controversy and debate among researchers. We summarize the main issues and discuss the solutions that have been presented. Even though the major issues have been dealt with extensively in the literature, there still exists quite a bit of confusion about how to analyze the data from such experiments. We discuss a number of the most frequently voiced objections. In particular, we discuss the

designs and the actual approach followed by many (if not most) researchers. In this paper I will try to analyze the reasons that may be responsible for this state of affairs. In doing so, I hope to be able to make it clear for an audience of nonstatisticians what the underlying

Canadian Journal of Experimental Psychology, 2003, 57:3, 141-151

participants, the mean observed difference between the conditions will be affected somewhat.

The third reason why the observed difference might be different is that the items that are used will be different (i.e., a valid replication will not necessarily use the same set of HF and LF words). Some words might be reacted to faster overall (the *main effect of items*); hence the mean difference between the HF

interaction term Treatment x List (within) does not exist for the case $p = 2$ (this interaction is then completely confounded with the Group main effect).

Since the same lists are used in all experimental

test in the item analysis does not test whether the effect is the same for all items but whether the population means are different, taking the variability between items into account (though ignoring the variability between subjects). In order to test whether the effect is the same for all items, one would have to test the Item x Treatment interaction effect (assuming such an effect does exist in the design). However, even if the Item x Treatment is significant, that says little or nothing about the difference between the population treatment means. Conversely, the fact that the treatment means are different does not imply that the effect holds for each and every item.

It is surprising (and somewhat disturbing) that such

show mo(arct lrvelmewhat ihs apnisininng such case). ns)Tj T* 1282eatmmhoisexa ime, weld havconstructedch set of datame f

fes
TTr

that in order to answer questions regarding the random or fixed character of experimental effects, one has to

across conditions is not feasible, the appropriate procedure is to calculate *minF'*. Second, if matching of items across conditions is possible, the optimal procedure would be to assume a blocked design. If matching is only possible at the set or list level, then one should use F_1 if one has sufficient confidence that the blocking was successful. If, however, there is reason to doubt the efficacy of the matching procedure, it might be better to use at least two lists (constructed according to the same matching procedure) for each treatment condition and to calculate the *F'* statistic as described by Clark (1973) with the average list score substituted for

xperitreatallowacroune bet-cace 15, 1018 WBd forTwbal Behavitedcon-'
12w (F')Tj /F7 1 1448.7836 0 TD 0,ed by247 1 0.Tj 3111373 0 TD -0.000(335-359.ed by) 33331. Genhe lizccordissulangu av pop capr

besoin d’effectuer des analyses d’items distinctes étant