Journal of Memory and Language **41,** 416–426 (1999)
Article ID jmla.1999.2650, available online at http://www.idealibrary.com on IDEAL®

How to Deal with "The Language-as-Fixed-Effect Fallacy": Common Misconceptions and Alternative Solutions

Jeroen G. W. Raaijmakers

University of Amsterdam, Amsterdam, The Netherlands

and

Joseph M. C. Schrijnemakers and Frans Gremmen

University of Nijmegen, Nijmegen, The Netherlands

Although Clark's (1973) critique of statistical procedures in language and memory studies (the "language-as-fixed-effect fallacy") has had a profound effect on the way such analyses have been carried out in the past 20 years, it seems that the exact nature of the problem and the proposed solution have not been understood very well. Many investigators seem to assume that generalization to both the subject population and the language as a whole is automatically ensured if separate subject (F_1) and item (F_2) analyses are performed and that the null hypothesis may safely be rejected if these F values are both significant. Such a procedure is, however, unfounded and not in accordance with the recommendations of Clark (1973). More importantly and contrary to current practice, in many cases there is no need to perform separate subject and item analyses since the traditional F_1 is the correct test statistic. In particular this is the case when item variability is experimentally controlled by matching or by counterbalancing.

Key Words: design; $minF'$; language; fixed effect; random effect.

Suppose that in a primed lexical decision experiment we want to investigate the effect of stimulus-onset asynchrony (SOA, the length of the interval between the onset of the prime and the onset of the target). To keep it simple, we use two levels of SOA. We start by selecting from some corpus a set of 40 related prime–target pairs. We divide this list randomly into two lists of 20 pairs, one for each SOA. In a within-subjects design, each subject is then presented both lists. Such a design is typical of many studies in the field of memory and language. How should these data be analyzed?

In a highly influential paper, Clark (1973) argued that the then-traditional way of analyzing such data (averaging the data for each subject over items within conditions and using these means in the ANOVA) was incorrect since it implicitly assumes that the materials variable (the individual word pairs) is a fixed factor and it does not take into account the fact that the items are sampled from a larger population of items. The major problem with this so-called "language-as-fixed-effect fallacy" is that it increases the probability of Type I errors, i.e., concluding that a treatment variable has an effect where in reality there is no such effect. The reason for this is not difficult to see: since some items are easier or are reacted to faster than others, the difference between the experimental conditions might be (partly) due to differences between the sets of items used in each of the conditions. Selecting language materials randomly or pseudorandomly leads to sampling variance that must be taken into account. Otherwise this variance will be confounded with the effect of the treatment variable. This problem had been previously discussed by Coleman (1964), but his paper did not get the attention it deserved and therefore did not have the impact that Clark's paper had.

The obvious solution to the language-as-

Address correspondence and reprint requests to Jeroen G. W. Raaijmakers, Department of Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands. E-mail: raaijmakers@psy.uva.nl.

0749-596X/99 $30.00

TABLE 1

Expected Mean-Squares for Repeated-Measurements ANOVA with Words Sampled Randomly

Source of variation	Label	*df*	Expected mean-squares
Treatment	A	$p - 1$	$\sigma^2_e + \sigma^2_{W(A)S} + q\sigma^2_{AS} + n\sigma^2_{W(A)} + nq\sigma^2_A$
Words (within Treatment)	W(A)	$p(q - 1)$	$\sigma^2_e + \sigma^2_{W(A)S} + n\sigma^2_{W(A)}$
Subjects	S	$n - 1$	$\sigma^2_e + \sigma^2_{W(A)S} + pq\sigma^2_S$
Treatment × Subjects	AS	$(p - 1)(n - 1)$	$\sigma^2_e + \sigma^2_{W(A)S} + q\sigma^2_{AS}$
Words × Subjects	W(A)S	$p(q - 1)(n - 1)$	$\sigma^2_e + \sigma^2_{W(A)S}$

Note. p = number of levels of the treatment variable; n = number of subjects; q = number of items. Words and Subjects are assumed to be random effects.

fixed-effect fallacy is to treat language materials as a random effect, as is the case with subjects. An effect is called random if the levels of that factor are sampled from some population. This is not a trivial aspect because whether an effect is treated as random or as fixed has consequences for the way in which the experimental effects should be tested.

In order to understand the problem, it may be helpful to consider the linear model that forms the basis for the ANOVA analysis. In the present case, the linear model is

$$X_{ijk} = \mu + \alpha_k + \beta_{j(k)} + \pi_i + \alpha\pi_{ik} + \pi\beta_{ij(k)} + \epsilon_{o(ijk)} \quad [1]$$

where μ = overall mean, α_k = main effect of experimental treatment k, $\beta_{j(k)}$ = main effect of word j (nested under treatment), π_i = main effect of subject i, $\alpha\pi_{ik}$ = the Treatment × Subject interaction, $\pi\beta_{ij(k)}$ = the Subject × Word interaction, and $\epsilon_{o(ijk)}$ = experimental error (in practice this term cannot be distinguished from the Subject × Word interaction, therefore these two terms are often combined into a single "residual" term). In the ANOVA, the variation in the experimental data is partitioned into independent sums-of-squares as shown in Table 1. Using the linear model of Eq. (1), it is possible to derive the expected values for the various sums-of-squares. These are shown in the rightmost column of Table 1.

In order to test for significance, an F ratio must be constructed in such a way that the expected value for the numerator is equal to the expected value of the denominator plus a term that reflects the effect to be tested. However, for the experimental design where both subjects and materials are treated as random-effect variables, the expected mean-squares for the various effects (see Table 1) are such that computation of a conventional F ratio is not possible. In order to see this, note that in order to test the treatment effect (A), i.e., the hypothesis $\sigma^2_A = 0$, we would need to construct an F ratio with the numerator equal to MS_A $(=\sigma^2_e + \sigma^2_{W(A)S} + q\sigma^2_{AS} + n\sigma^2_{W(A)} + nq\sigma^2_A)$ and in the denominator a term with expected mean-squares equal to $\sigma^2_e + \sigma^2_{W(A)S} + q\sigma^2_{AS} + n\sigma^2_{W(A)}$.[1] As can be seen in Table 1, no such term exists. The traditional solution to such problems is to compute a quasi F ratio, F':

$$F' = \frac{MS_A + MS_{W(A)S}}{MS_{AS} + MS_{W(A)}}. \quad [2]$$

F' has an approximate F-distribution with degrees of freedom for the numerator and the denominator given by

$$df = (MS_1 + MS_2)^2/(MS_1^2/df_1 + MS_2^2/df_2), \quad [3]$$

where MS_1 and MS_2 are the two mean-squares in the numerator or the denominator and df_1 and df_2 are the corresponding degrees of freedom (see Clark, 1973, p. 338). The rationale behind

[1] For simplicity, the notation σ^2_A is used, irrespective of whether the effect A is fixed or random.