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. © 1999 Academic Press

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

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

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':

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

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

$$df = (MS_1 + MS_2)^2/(MS^2_1/df_1 + MS^2_2/df_2), \qquad [3]$$

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.

TABLE 2

Simulated Data for Repeated-Measurements ANOVA with Words Sampled Randomly

	Short SOA				Long SOA			
Subject	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
1	546	567	547	566	554	545	594	522
2	566	556	538	566	512	523	569	524
3	567	598	568	584	536	539	589	521
4	556	565	536	550	516	522	560	486
5	595	609	585	588	578	540	615	546
6	569	578	560	583	501	535	568	514
7	527	554	535	527	480	467	540	473
8	551	575	558	556	588	563	631	558

the use of F' becomes evident when the expected values for the mean-squares are substituted in the equation:

$$E(F') \approx \frac{E(MS_{\mathrm{A}}) + E(MS_{\mathrm{W(A)S}})}{E(MS_{\mathrm{AS}}) + E(MS_{\mathrm{W(A)}})} = \frac{2\sigma^2_{\mathrm{e}} + 2\sigma^2_{\mathrm{W(A)S}} + \mathrm{q}\sigma^2_{\mathrm{AS}} + \mathrm{n}\sigma^2_{\mathrm{W(A)}} + \mathrm{nq}\sigma^2_{\mathrm{A}}}{2\sigma^2_{\mathrm{e}} + 2\sigma^2_{\mathrm{W(A)S}} + \mathrm{q}\sigma^2_{\mathrm{AS}} + \mathrm{n}\sigma^2_{\mathrm{W(A)}}}. \quad [4]$$

As can be seen from this equation, F' has the structure of regular F statistics, i.e., the numerator is equal to the denominator plus one extra term corresponding to the effect to be tested. However, since it is not based on the computation of independent sums-of-squares, it is not a true F statistic and only approximately distributed as F, although it is the best approach available given that a true F statistic is not available.

Table 2 gives a numerical example with simulated data (example data have been included for most designs discussed in this article; the purpose of these examples is not to demonstrate a particular point but primarily to enable the interested reader to verify the results by carrying out the appropriate analyses using his/her favorite statistical package). In the present example, eight subjects are each tested under two conditions (a short and long SOA, respectively). There are eight items, four of which are randomly assigned to each of the two conditions. The data were generated using a model in which there was no real effect of condition. Table 3 gives the ANOVA table corresponding to these data. Application of Eq. (2) gives $F'(1,9) = 1.70$.

In practice, however, F' will be difficult to compute due to missing data (e.g., error responses) and limitations in the size of the ANOVA designs in most statistical packages (especially if a program based on the General Linear Model is used). It is then easier to compute its lower bound *minF'*, using the F values of separate subject and item analyses, usually referred to as F_1 and F_2, respectively. In a *subject analysis* each data point in a cell of the design is computed by collapsing over items, whereas in an *item analysis* data points are computed by collapsing over subjects. Although *minF'* is linked to an analysis that treats subjects and language materials (items) as random effects in a *single* ANOVA model, this statistic

TABLE 3

ANOVA Summary Table for Example Data for Table 2

Source of variation	*df*	Mean-square
Treatment	1	8032.6
Words (within Treatments)	6	3695.7
Subjects	7	3750.2
Treatment × Subjects	7	1083.8
Words × Subjects	42	100.2

can be computed by the F values from separate subject and item analyses. As shown by Clark (1973):

$$minF' = \frac{MS_{\mathrm{A}}}{MS_{\mathrm{AS}} + MS_{\mathrm{W(A)}}} = \frac{F_1 F_2}{F_1 + F_2}. \qquad [5]$$

For the data in Table 2, $F_1(1,7) = 7.41$ and $F_2(1,6) = 2.17$, hence $minF'(1,10) = 1.68$. It is evident that using F_1 would lead to an incorrect conclusion. In this example, F_2 does much better and $minF'$ is actually quite close to F'. These data reiterate the point made by Clark (1973) about the bias that would be present if F_1 was used to test the difference between the conditions.

Clark's paper was highly influential and it is now customary (especially among language researchers) to routinely run both an item and a subject analysis. But it appears that there has been some misconception with respect to the nature of the problem and the solution proposed by Clark (1973). Many researchers have been testing their treatment effects on the basis of separate subject and item analyses and have rejected the null hypothesis if both analyses showed significant F values. However, this procedure, which will hereafter be denoted as the $F_1 \times F_2$ criterion, is *not* equivalent to the $minF'$ solution and leads to positive bias (a higher α than the nominal α) if item variance is not controlled for, as a theoretical analysis shows and as Forster and Dickinson (1976) demonstrate by Monte Carlo simulations. Of course asserting a difference when either F_1 or F_2 is significant would result in an even greater bias.

In this article we first review Clark's solution and show that the $F_1 \times F_2$ criterion, although widely used, leads to positive bias. Next, we discuss alternatives to the $minF'$ approach and consider the effects of commonly used variations in the exact nature of the design (such as matching of items and counterbalancing of lists) that affect the outcome of the analysis. We hope to convince the reader that it is necessary to take the details of the experimental design into account before deciding on the particular ANOVA to be performed.

CURRENT PRACTICE: THE $F_1 \times F_2$ FALLACY

Although there was some controversy in the late 1970s regarding the necessity and appropriateness of treating items as a random factor in the analysis of experiments with language materials (see Cohen, 1976; Smith, 1976; Wike & Church, 1976; see also Clark, 1976), by the early 1980s the issue was more or less settled and researchers started to routinely perform both subject and item analyses. However, many researchers seem to believe that the subject analysis (F_1) makes it possible to test for reliability of the effect over subjects and that the item analysis (F_2) makes it possible to test for reliability of the effect over items. Hence, if both F statistics are significant, it should (according to this reasoning) be the case that the effect is reliable over both subjects and items.

However, this is incorrect since in the standard design considered by Clark (1973) *both* F_1 and F_2 will be biased when subjects and items are sampled randomly. To see this, note that if F_1 and F_2 are equal, $F_1 = F_2 = F$, hence $minF' = F/2$. Thus, both F_1 and F_2 could be significant, while $minF'$ would not. If this happens, many researchers seem to be hesitant to accept that the effect is not significant. An example may be found in Katz (1989, p. 492). After obtaining an F_1 of 10.8 and an F_2 of 5.44 (both p's $< .05$), Katz reluctantly concludes: "The effect of concreteness was marginally significant when the overly conservative $minF$ test was computed; $minF(1,44) = 3.62$, $p < .10$." Most researchers today do not even compute or report the value of $minF'$. In some cases a rather curious mixed approach is used. For example, Seidenberg et al. (1984, p. 386) report: "Min F' statistics were calculated, and are reported when they were significant; otherwise, the significant F statistics for the subject and item analyses are reported."

Sometimes this procedure is justified by the argument that the $minF'$ procedure is a too conservative test (see the quote above) and that this $F_1 \times F_2$ procedure avoids both this bias in

$minF'$ as well as the bias in F_1. For example, Smith (1976) and Wike and Church (1976) criticized the use of F' (and $minF'$) as being an unduly conservative test. The power of the test based on F' depends on a number of factors such as the structure of F' (the terms included in its calculation), the size of the error variance, the number of the degrees of freedom, and the number of levels of the treatment variable. However, Monte Carlo simulations with F' as in Eq. (2) have demonstrated that it is a good approximation to the normal F statistic (Davenport & Webster, 1973; Forster & Dickinson, 1976). Furthermore, as Forster and Dickinson (1976) have shown, $minF'$ is a good estimate of F', and both statistics are not unduly conservative, given that $\sigma^2_{\mathrm{W(A)}}$ and σ^2_{AS}, the variance components expressing item and subject variability, are not too small (relative to $\sigma^2_{\mathrm{W(A)S}}$). In most experiments this is likely to be the case.

In addition, Wike and Church (1976) commented that although F' has an approximate F distribution, little is known about the characteristics of its distribution. In most psycholinguistic and semantic memory tasks dependent variables such as reaction times are not normally distributed. The conventional F test is robust against violations of homogeneity of variance and normality of the distribution of the dependent variable. However, Santa, Miller, and Shaw (1979) demonstrated that the F' and $minF'$ are also robust against violations of homogeneity and normality. They showed by means of Monte Carlo simulations that with heterogeneous treatment group variances and with five types of error distributions (normal, exponential, log-uniform, binary, and log-normal) the F' and $minF'$ have real alpha values that are near the nominal alpha value of .05. Only when the variance components $\sigma^2_{\mathrm{W(A)}}$ and σ^2_{AS} are small do both statistics tend to be conservative.

Thus, there is no justification for the assertion that the $minF'$ procedure advocated by Clark (1973) is too conservative. Hence, the argument that the $F_1 \times F_2$ procedure may be justified because of the conservative nature of $minF'$ is incorrect. Rather, the situation is the other way around. As Forster and Dickinson (1976) have shown by means of Monte Carlo simulations, the $F_1 \times F_2$ procedure has a larger error rate than .05 although of course not as large as separate F_1 or F_2 analyses.

To give some indication of this "$F_1 \times F_2$ fallacy", we screened Volumes 32–37 (1993–1997) of the *Journal of Memory and Language* and counted how many of the published papers reported both F_1 and F_2 in at least one experiment. In these volumes there were a total of 220 papers, of which 120 reported F_1 and F_2 without mentioning any $minF'$ values. In only 4 papers were F_1, F_2, and $minF'$ values reported. Thus, these statistics clearly show that the $F_1 \times F_2$ fallacy is quite widespread.[2]

Widespread as it may be, this fallacy did not become common practice immediately after the appearance of Clark's paper. Initially, researchers based their conclusions on the outcome of the $minF'$ test and often did not even report the F_1 and F_2 statistics. Today, the situation is completely the opposite. To demonstrate this, we first counted all papers between 1974 and 1997 of the *Journal of Verbal Learning and Verbal Behavior/Journal of Memory and Language* that reported $minF'$ and/or F_1 and F_2 and then computed the proportion of those papers that reported $minF'$ and not just F_1 and F_2. These proportions are plotted in Fig. 1. The results are clear: there is a steady change from a situation in which the $F_1 \times F_2$ criterion is never used to a situation in which the $minF'$ criterion is almost never used.

Of course, the conclusions that were based on the $F_1 \times F_2$ testing procedure in the screened papers need not be incorrect. As we show below, if the researchers experimentally controlled for item variability, the use of F_1 by itself might have been the correct procedure. In that case, the use of $F_1 \times F_2$ would only have been a more conservative procedure and all significant results would remain significant (although some results that were reported as not

[2] It is of some interest that the use of item analyses and $minF'$ seems to be restricted to those analyses in which the dependent variable is reaction time, although from a statistical point of view there is no reason why the "language-as-fixed-effect" issue should not be relevant when accuracy measures are analyzed.

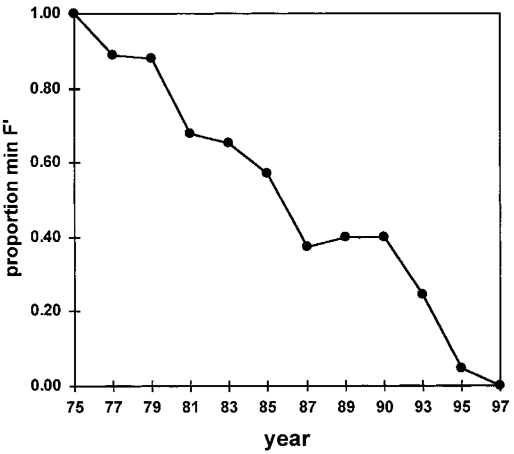

FIG. 1. The proportion of papers that report $minF'$ of all papers that report F_1 and F_2 (based on a count of all papers in *JVLVB*/*JML* between 1974 and 1997). Data are grouped in 2-year intervals.

significant might in fact be significant). However, if in these studies subjects and items were in fact sampled (quasi-)randomly the use of $minF'$ might have led to different conclusions.

MATCHING OF ITEMS

Error variance introduced by random or pseudorandom selection of items can be controlled by statistical or experimental procedures. Statistical control can be achieved by adding an item variance component to the ANOVA model, and hypothesis testing is then based on the computation of F' or its lower bound $minF'$. This was the solution proposed by Clark. Item variance can, however, also be controlled by experimental procedures such as matching stimulus materials in different treatment groups with respect to variables that correlate highly with the dependent variable. This seems to be the preferred approach in current research. It is rarely the case that investigators select their stimulus materials in a truly random fashion. Normally, items are carefully selected and balanced on important variables that correlate with the response measure. Of course, if balancing is used to control item variance it will replace the requirement to control by statistical procedures, i.e., applying the random effect model (F'). In that case item variability would not affect the differences between the treatment conditions and it would be best to perform a subject (F_1) analysis, where subjects form the only random variable. Wickens and Keppel (1983) showed that if item variance is controlled in this manner, the bias in F_1 is indeed greatly reduced. Moreover, if the blocking factor is ignored in the analysis, it is best to perform a F_1 analysis because in that case the usage of F' or $minF'$ leads to a considerable reduction in power (see Wickens & Keppel, 1983, p. 307). Of course, if it were possible to do the full analysis (including subjects as well as blocks), that would be the preferred analysis. However, as explained earlier, this is rarely possible.

To illustrate these points, we will take a closer look at the ideal case in which this type of blocking or matching captures all of the systematic variability between items. That is, the two (or more if there are more treatment conditions) items within a block are perfectly matched. The various blocks are still assumed to be sampled randomly from a larger population of blocks. The major difference in such a design is that the blocks factor will be *crossed* with treatments instead of being nested under treatments as was the case when items are randomly sampled. To make it easier to understand the nature of this design, we constructed a small set of simulated data in which there are again eight subjects, each tested under two conditions (see Table 4). Suppose that we are able to select pairs of items in such a way that they are matched on the most important item variables that affect the lexical decision times. Hence, there will be four pairs of matched items or blocks. Within each block, one item is assigned to each of the two experimental conditions. Note that both items of a given pair of matched items have been given the same block label in order to emphasize the blocking. The data were again generated using a model in which there was no real effect of condition.

Table 5 gives the expected mean-squares for such a design. This case is similar to the traditional case in that here too there is no simple F statistic to test the significance of the treatment effect. A quasi-F ratio that may be used to

TABLE 4

Simulated Data for Repeated-Measurements ANOVA with Matched Items

	Short SOA				Long SOA			
Subject	Block 1	Block 2	Block 3	Block 4	Block 1	Block 2	Block 3	Block 4
1	493	519	513	542	499	525	502	557
2	562	552	565	591	544	536	533	563
3	519	558	555	567	575	582	551	587
4	518	523	514	563	523	565	539	597
5	567	562	577	595	521	563	559	575
6	520	534	527	568	512	541	531	559
7	516	544	513	575	555	569	550	601
8	525	528	528	559	551	542	529	578

evaluate the significance of the treatment effect, is given by

$$F' = \frac{MS_{\mathrm{A}} + MS_{\mathrm{ABS}}}{MS_{\mathrm{AS}} + MS_{\mathrm{AB}}}. \qquad [6]$$

Wickens and Keppel (1983) showed that in such a design with blocking of materials, the bias in the F ratio from the standard subject analysis (F_1) is greatly reduced. To see this more clearly, note that

$$E(F_1)$$

$$\approx \frac{E(MS_{\mathrm{A}})}{E(MS_{\mathrm{AS}})} \qquad [7]$$

$$= \frac{\sigma_{\mathrm{e}}^2 + \sigma_{\mathrm{ABS}}^2 + \mathrm{q}\sigma_{\mathrm{AS}}^2 + \mathrm{n}\sigma_{\mathrm{AB}}^2 + \mathrm{nq}\sigma_{\mathrm{A}}^2}{\sigma_{\mathrm{e}}^2 + \sigma_{\mathrm{ABS}}^2 + \mathrm{q}\sigma_{\mathrm{AS}}^2}.$$

Hence the bias in F_1 is now a function of σ_{AB}^2, the interaction between blocks and treatments, and this will usually be smaller than $\sigma_{\mathrm{W(A)}}^2$, the variability of items within treatments that is responsible for the bias in the case where items are sampled randomly (i.e., not matched).

Table 6 gives the full ANOVA table corresponding to the example data of Table 4. Applying Eq. (6) gives $F'(1{,}8) = 0.87$. For these data, $F_1(1{,}7) = 0.86$ and $F_2(1{,}3) = 7.19$ (if the matching is taken into account, i.e., if a repeated-measures design is used in the item analysis, as would be appropriate), hence $minF'(1{,}3) = 0.77$. If the matching is not taken into account, $F_2(1{,}6) = 0.27$, hence $minF'(1{,}10) = 0.20$. It is evident that if the matching is taken into account both $minF'$ and F_1 give a good approximation to the “true” F'. If the matching is not taken into account, F_2 is quite a bit smaller and $minF'$ underestimates the

TABLE 5

Expected Mean-Squares for Repeated-Measurements ANOVA with Blocks or Matched Items Crossed with Treatments

Source of variation	*df*	Expected mean-squares
A (Treatment)	$p - 1$	$\sigma_{\mathrm{e}}^2 + \sigma_{\mathrm{ABS}}^2 + \mathrm{q}\sigma_{\mathrm{AS}}^2 + \mathrm{n}\sigma_{\mathrm{AB}}^2 + \mathrm{nq}\sigma_{\mathrm{A}}^2$
B (blocks)	$q - 1$	$\sigma_{\mathrm{e}}^2 + \mathrm{p}\sigma_{\mathrm{BS}}^2 + \mathrm{np}\sigma_{\mathrm{B}}^2$
S (Subjects)	$n - 1$	$\sigma_{\mathrm{e}}^2 + \mathrm{p}\sigma_{\mathrm{BS}}^2 + \mathrm{pq}\sigma_{\mathrm{S}}^2$
A × B	$(p - 1)(q - 1)$	$\sigma_{\mathrm{e}}^2 + \sigma_{\mathrm{ABS}}^2 + \mathrm{n}\sigma_{\mathrm{AB}}^2$
A × S	$(p - 1)(n - 1)$	$\sigma_{\mathrm{e}}^2 + \sigma_{\mathrm{ABS}}^2 + \mathrm{q}\sigma_{\mathrm{AS}}^2$
B × S	$(q - 1)(n - 1)$	$\sigma_{\mathrm{e}}^2 + \mathrm{p}\sigma_{\mathrm{BS}}^2$
A × B × S	$(p - 1)(q - 1)(n - 1)$	$\sigma_{\mathrm{e}}^2 + \sigma_{\mathrm{ABS}}^2$

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

TABLE 6

ANOVA Summary Table for Example Data of Table 4

Source of variation	*df*	Mean-square
A (Treatment)	1	770.1
B (blocks)	3	5661.8
S (Subjects)	7	1822.6
A × B	3	107.1
A × S	7	893.5
B × S	21	143.8
A × B × S	21	102.1

correct value of F'. As we mentioned before, Wickens and Keppel (1983) showed that this is a general finding in this type of design. Thus, if it is not possible (because of missing data) to do the full analysis that includes the blocking factor (leading to F'), it is better to simply use F_1 rather than *min*F', especially when the matching of the items is not taken into account in the item analysis used to determine F_2.

COUNTERBALANCED DESIGNS

In many cases, however, better ways of controlling item variability are possible. One such approach involves the case where items are sampled randomly for each subject separately. In this case each subject receives a different set of words under each of the treatment levels. This case was briefly mentioned by Clark (1973, p. 348) as one where the traditional analysis (F_1) is correct (see also Winer, 1971, p. 365). In such a design where Items are nested within Subjects and Treatments, the treatment effect may be tested against the Treatment × Subjects interaction, which is equivalent to the regular F_1 test when the data are collapsed over items.

An alternative approach that is frequently used in memory research is the use of counterbalanced lists. In such a design, one group of subjects receives List 1 in condition 1 and List 2 in condition 2, and a second group of subjects receives List 2 in condition 1 and List 1 in condition 2. In this design, the between-groups variability is confounded with (part of) the interaction between list and treatment. However, and this is of course the rationale for using such a design, the mean difference between the treatment conditions (and hence the treatment effect) is not affected by any difference that might exist between the lists.

Table 7 gives a numerical example with three experimental conditions and three lists of four items each. Hence there are three groups of subjects and the assignment of lists to conditions is counterbalanced across groups. As before, it is normally not possible to analyze such a complete design and the experimenter will have to average the scores for the four items within each condition. These averages are given in Table 8. How should such data be analyzed, taking into account that the factor Lists should be a random effect?

In order to answer this question, we will take a closer look at the expected mean-squares for this design (see Table 9). This type of design is discussed by Winer (1971, pp. 712, 716) and Kirk (1982, p. 648), although they treated the factor corresponding to Lists as fixed. Kirk refers to this design as a Latin Square Confounded Factorial design (LSCF). The ANOVA model for this design is as follows:

$$X_{\mathrm{ijm(t)}} = \mu + \theta_{\mathrm{t}} + \pi_{\mathrm{m(t)}} + \alpha_{\mathrm{i}} + \beta_{\mathrm{j}} + \alpha\beta'_{\mathrm{ij}} + \epsilon_{\mathrm{ijm(t)}}, \qquad [8]$$

where μ = overall mean; θ_{t} = effect of group t (= the between-component of the Treatment × List interaction); $\pi_{\mathrm{m(t)}}$ = effect of subject m (nested within group t), α_{i} = effect of the experimental treatment i; β_{j} = effect of list j; $\alpha\beta'_{\mathrm{ij}}$ = the within-component of the Treatment × List interaction; and $\epsilon_{\mathrm{ijm(t)}}$ = experimental error (a residual term equivalent to the interaction between Treatment, List, and Subjects plus “real” error; this term might be further decomposed but this would not affect the results). Due to the nature of this design (each group receives only p of the $p \times p$ combinations of Treatment and List), the interaction between Treatment and List is divided into two components, one between subjects and one within subjects.

In the ANOVA model it is assumed that Group, Subjects within Groups, as well as List

TABLE 7

Simulated Data for Design Using Counterbalanced Lists

		Short SOA				Medium SOA				Long SOA			
Group	Subject	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12
1	1	532	508	522	482	468	496	544	547	475	522	502	484
	2	542	516	545	483	509	519	588	583	499	535	535	486
	3	615	584	595	560	542	591	630	617	543	606	560	545
	4	547	553	584	535	514	555	591	606	538	565	546	527
		Item 9	Item 10	Item 11	Item 12	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
2	5	553	598	581	551	619	576	606	561	548	590	614	631
	6	464	502	485	451	484	479	499	471	447	486	514	523
	7	481	511	492	472	531	506	542	475	471	510	539	556
	8	541	588	551	533	582	556	589	515	538	545	601	576
		Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 1	Item 2	Item 3	Item 4
3	9	482	530	571	563	501	561	500	506	543	539	558	497
	10	559	570	632	639	551	592	572	561	617	587	616	549
	11	462	497	546	538	487	546	491	470	529	508	525	473
	12	460	463	511	528	457	506	487	453	498	479	512	443

are random factors (Group is random since it corresponds to an interaction between a fixed and a random effect). That is, it is assumed that the lists are based on a random sample of words from a larger population of words. Table 9 gives the expected mean-squares for this design under these assumptions. Note that the interaction term Treatment × List (within) does not exist for the case $p = 2$ (this interaction is then completely confounded with the Group main effect).

As can be seen from Table 9, in order to test the treatment effect, the treatment mean-square should be tested against the Treatment × List (within) mean-square. If the F test for the Treatment × List (within) interaction effect is not significant by a conservative criterion ($\alpha = .25$), this mean-square may be pooled with the error (residual) mean-square, giving a much more powerful test for the treatment effect. In the special case where $p = 2$, the treatment effect is always tested against the error mean-square. Hence, in all of these cases there is no necessity to run two analyses, one over subjects and one over items. The subject analysis (i.e., the analysis in which all items from a single list are averaged) will give all the information that is required to test the significance of the treatment effect. Moreover, there is no necessity to compute a quasi-F ratio: regular F-ratios will be correct.

This design was also discussed by Pollatsek and Well (1995, Table 4), except that they denoted the main effect of List as the Groups × Treatment interaction. However, as mentioned above, these two effects are equivalent in the $p = 2$ case. In the case $p > 2$, the Groups × Treatment interaction consists of two terms, namely the main effect of List and the within-part of the Treatment × List interaction. In order to separate these effects, two ANOVA's should be run, the traditional one without the List effect but with the Groups × Treatment interaction and a second one with the List effect but without the Groups × Treatment interaction. The latter analysis gives the correct value for the List sums-of-squares, and subtracting this from the sums-of-squares for the Groups × Treatment interaction gives the correct value for the within-part of the Treatment × List interac-

TABLE 8

Data from Table 7 Collapsed over Items

Group	Subject	Short SOA List 1	Medium SOA List 2	Long SOA List 3
1	1	511	514	496
	2	522	550	514
	3	588	595	563
	4	554	567	544
		List 3	List 1	List 2
2	5	571	591	596
	6	476	483	492
	7	489	514	519
	8	553	560	565
		List 2	List 3	List 1
3	9	536	517	534
	10	600	569	592
	11	511	498	509
	12	490	476	483

tion. This partitioning of the Groups × Treatment interaction was also briefly discussed by Pollatsek and Well (1995, Appendix A, especially Table A2), although the expected mean-squares that they present apply to the case that the List factor is assumed fixed rather than random. Contrary to the suggestion of Pollatsek and Well (1995), however, it is not required to do separate analyses over subjects and items in order to test the effect of the treatment factor. The expected mean-squares, under the assumption that List is a random effect, show that the Treatment effect can always be tested directly using the mean-squares obtained from the standard subject analysis (averaging over items).

In Table 10 we present the ANOVA summary table for the data presented in Table 8. As explained above, the Treatment × List Sums-of-Squares was obtained by subtracting the List main effect Sums-of-Squares (3106.2) from the Group × Treatment Sums-of-Squares (3152.3). If we test the Treatment main effect against the Treatment × Lists interaction, the resulting F ratio equals $F(2,2) = 1.116$. However, since the Treatment × Lists interaction is not significant [$F(2,18) = 0.786$], a more powerful test may be obtained by pooling this interaction and the error Sums-of-Squares and testing the treatment effect against this pooled error. Note that this pooled error Sums-of-Squares may be obtained directly from the analysis that includes the List main effect but not the Group × Treatment interaction effect. This gives an error term for the F test that is based on 20 degrees of freedom instead of just 2. In the present example, the resulting F value is 0.896.

CONCLUSION

There are two important conclusions that we draw from these analyses. The first is that many language researchers are applying statistical procedures that do not match the details of the actual design that they are using. In many cases the design does not require separate analyses over subjects and items, yet such analyses are routinely run, without taking into account that this procedure was originally introduced for a

TABLE 9

Expected Mean-Squares for Repeated-Measurements ANOVA with Counterbalanced Lists

Source of variation	*df*	Expected mean-squares
G (groups) (=A × L between)	$p - 1$	$\sigma_e^2 + p\sigma_{S(G)}^2 + np\sigma_G^2$
S(G)	$p(n - 1)$	$\sigma_e^2 + p\sigma_{S(G)}^2$
A	$p - 1$	$\sigma_e^2 + n\sigma_{AL}^2 + np\sigma_A^2$
L (lists)	$p - 1$	$\sigma_e^2 + np\sigma_L^2$
A × L (within)	$(p - 1)(p - 2)$	$\sigma_e^2 + n\sigma_{AL}^2$
Residual	$p(n - 1)(p - 1)$	σ_e^2

Note. p = number of groups = number of levels of the treatment variable = number of lists; n = number of subjects within each group. Lists and Subjects are assumed to be random effects.

TABLE 10

ANOVA Summary Table for Example Data of Table 8

Source of variation	*SS*	*df*	*MS*	*F*
G (groups) (=A × L between)	1720.2	2	860.08	0.164
S(G)	47206.2	9	5245.13	178.527
A	51.5	2	25.75	1.116
L (lists)	3106.2	2	1553.08	52.862
A × L (within)	46.2	2	23.08	0.786
Error	528.8	18	29.38	
Error (pooled)	575.0	20	28.75	

very specific design, namely a design where the items are nested under the treatment variable. If this is not in fact the case, e.g., when the materials have been matched on a number of variables or when the lists are counterbalanced over different groups of subjects, there is no need to compute (min) F' and the simple subject analysis (averaging over items) will be correct. The second conclusion is that the practice of running both a subject and an item analysis and of using the $F_1 \times F_2$ criterion is both widespread as well as without any foundation. Either F_1 is correct or it is incorrect. In the latter case, (min) F' is the correct statistic to compute.

REFERENCES

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.

Clark, H. H. (1976). Reply to Wike and Church. *Journal of Verbal Learning and Verbal Behavior,* **15,** 257–261.

Cohen, J. (1976). Random means random. *Journal of Verbal Learning and Verbal Behavior,* **15,** 261–262.

Coleman, E. B. (1964). Generalizing to a language population. *Psychological Reports,* **14,** 219–226.

Davenport, J. M., & Webster, J. T. (1973). A comparison of some approximate *F*-tests. *Technometrics,* **15,** 779–789.

Forster, K. I., & Dickinson, R. G. (1976). More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F_1, F_2, F', and $minF'$. *Journal of Verbal Learning and Verbal Behavior,* **15,** 135–142.

Katz, A. N. (1989). On choosing the vehicles of metaphors: Referential concreteness, semantic distances, and individual differences. *Journal of Memory and Language,* **28,** 486–499.

Kirk, R. E. (1982). *Experimental design: procedures for the behavioral sciences.* Pacific Grove: Brooks/Cole.

Pollatsek, A., & Well, A. D. (1995). On the use of counterbalanced designs in cognitive research: A suggestion for a better and more powerful analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **21,** 785–794.

Santa, J. L., Miller, J. J., & Shaw, M. L. (1979). Using quasi *F* to prevent alpha inflation due to stimulus variation. *Psychological Bulletin,* **86,** 37–46.

Seidenberg, M. S., Waters, G. S., Barnes, M. A., & Tanenhaus, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning and Verbal Behavior,* **23,** 383–404.

Smith, J. E. K. (1976). The assuming-will-make-it-so fallacy. *Journal of Verbal Learning and Verbal Behavior,* **15,** 262–263.

Wickens, T. D., & Keppel, G. (1983). On the choice of design and of test statistic in the analysis of experiments with sampled materials. *Journal of Verbal Learning and Verbal Behavior,* **22,** 296–309.

Wike, E. L., & Church, J. D. (1976). Comments on Clark's "The language-as-fixed-effect fallacy." *Journal of Verbal Learning and Verbal Behavior,* **15,** 249–255.

Winer, B. J. (1971). *Statistical principles in experimental design.* New York: McGraw–Hill.

(Received September 22, 1997)
(Revision received February 19, 1999)