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

Jeroen G. W. Raaijmakers, University of Amsterdam

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 issue of what happens if in a counterbalanced design only some of the items show the treatment effect. Finally, a possible solution is discussed for the case where only partial matching of items between conditions is possible.

One of the most intriguing controversies in the use of statistics in behavioural science concerns what is known as the "language-as-fixed-effect fallacy." This controversy refers to the statistical problems that occur when in an experiment the items that are presented to a subject are drawn from a population of items and the researcher wishes to draw conclusions that are valid not just for the particular sample of items used in the experiment but for the population of items from which that sample was drawn. The controversy is intriguing because there is a remarkable discrepancy between the published recommendations for how to analyze such 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 issues are that should be considered when evaluating a particular approach to this problem.

As its name already suggests, the issue becomes relevant whenever a researcher designs an experiment in which natural language materials are used as stimuli. This also explains why the issue has been most strongly debated by researchers in psycholinguistics, although the problem is certainly not unique to language research (memory researchers, for example, hardly ever pay attention to the problem even though there are quite a few cases where that would be warranted). Moreover, exactly the same problem of generalization to a population of stimulus items occurs in research using other types of stimuli such as pictures.

To make the issue more concrete, suppose that a researcher wants to test the effects of language frequency on lexical decision times. In order to do so, two sets of 20 items each are selected, one of high-frequency nouns (HF) and one of low-frequency nouns (LF). The two sets of items are presented in a mixed list together with a set of nonwords to a group of, say, 40 participants. For the present purposes we will focus on the lexical decision times for the words and we will disregard the nonword data.

Let us assume that the researcher observes a mean difference of 30 ms between the HF and LF conditions. Obviously, such a difference cannot be taken at face value but will have to be evaluated in light of the variability of this outcome that will inevitably occur when the experiment would be repeated. In order to get some idea of the magnitude of this variability, we have to consider which aspects might be different in an independent replication of the experiment. In this experiment, there are four reasons why the outcome might be different. The first and simplest one is of course the inherent variability or *random error* that is due to the fact that even when everything is kept the same (same design, same participants, same items), the outcome will always be somewhat different.

The second one is that in a replication the participants will be different. Since each participant gets both conditions, any overall difference between participants (i.e., some participants will be faster overall) will not affect the difference between the mean RT in the two conditions. However, some participants will have a somewhat greater difference between the two conditions (i.e., the *Subjects x Treatment interaction effect*). Hence, when the experiment is replicated with new

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	$\sigma_e^2 + \sigma_{W(A)S}^2 + q\sigma_{AS}^2 + n\sigma_{W(A)}^2 + nq\sigma_A^2$
Words (within Treatment)	W(A)	$p(q$-1)	$\sigma_e^2 + \sigma_{W(A)S}^2 + n\sigma_{W(A)}^2$
Subjects	S	n-1	$\sigma_e^2 + \sigma_{W(A)S}^2 + pq\sigma_S^2$
Treatment x Subjects	AS	(p-1)(n-1)	$\sigma_e^2 + \sigma_{W(A)S}^2 + q\sigma_{AS}^2$
Words x Subjects	W(A)S	$p(q$-1)(n-1)	$\sigma_e^2 + \sigma_{W(A)S}^2$

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.

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 and LF conditions will vary when a new set of items is used. In addition to the main effect of items, the difference between the HF and LF conditions will also be affected by the fact that some subjects will react relatively faster to some items than others, the *Item x Subjects interaction effect* (the Item x Treatment interaction effect is not defined in this design since items are nested under treatments). Note, however, that in this design this effect will be completely confounded with the random error component.

The above analysis may also be formulated using the standard linear model that is used in the analysis of variance approach. In the present case, the linear model is

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

where μ = overall mean, α_k = main effect of experimental treatment k, $\beta_{j(k)}$ = main effect of item j (nested under treatment condition k), π_i = main effect of subject i, $\alpha\pi_{ik}$ = the Treatment x Subject interaction, $\pi\beta_{ij(k)}$ = the Subject x Item interaction, and $\varepsilon_{o(ijk)}$ = experimental error (the dummy subscript o is conventionally used to indicate that the experimental error is nested within the individual observation; as mentioned above, this term cannot be distinguished from the Subject x Item interaction; therefore these two terms are often combined into a single "residual" term). In the ANOVA approach, the differences between the experimental conditions are expressed as sums-of-squares. In the present example, the sums-of-squares corresponding to the difference between the HF and LF items is given by SS_A and is proportional to the squared difference between the treatment means. Similar sums-of-squares may be defined for other averages that might be calculated (i.e. the differences between the items, the subjects etc.). Thus, the variation in the experimental data is partitioned into sums-of-squares as shown in Table 1.

When we perform a statistical significance test for the difference between the treatment means, we are essentially determining whether the observed difference is larger than might be expected on the basis of the other sources of variation that contribute to the variability between the treatment means. Which sources of variation are contributing to the difference between the treatment means may be determined by calculating the expected values for the sums of squares (or rather the mean squares, i.e., the corresponding variances). For most experimental designs, these expected values are listed in standard textbooks on ANOVA or may be obtained from simple algorithms (such as the Cornfield-Tukey algorithm). For the linear model of Equation 1, the expected mean squares are listed in the rightmost column of Table 1.

As can be seen in Table 1, the expected value for the treatment mean square is equal to $\sigma_e^2 + \sigma_{w(A)S}^2 + q\sigma_{AS}^2 + n\sigma_{w(A)}^2 + nq\sigma_A^2$, which corresponds to the sources of variation discussed earlier plus the "real" treatment effect (σ_A^2).[1] That is, σ_e^2 corresponds to the random error variation, σ_{AS}^2 to the Subjects x Treatment interaction effect, $\sigma_{w(A)}^2$ to the main effect of items, and $\sigma_{w(A)S}^2$ corresponds to the Item x Subjects interaction effect.

In this derivation two assumptions are made. First, replications of the experiment will use different random samples of HF and LF items in the two experimental conditions. This is essentially what it means when it is said that Items is a *random* factor. If exactly the same items would be used in every possible replica-

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

tion, the factor would be called *fixed*. If that was the case, the term corresponding to the item variation would drop from the formula for the expected mean square for the treatment effect. This follows from the observation that the items no longer make a contribution to the variation in the difference between the treatment means in independent replications of the experiment. The second assumption is that the variation in the difference in the item means *between conditions* may be estimated from the variation *within the conditions*. That is, the variability in the item means for the two conditions is directly related to the variability of the items within each condition. This is usually expressed by the assumption that the items within each condition are sampled *randomly* and *independently* from populations having identical distributions (i.e., each distribution has the same variance, $\sigma^2_{W(A)}$). As we will see later, there are designs where this latter assumption would not be correct, for example, when the items are selected in such a way as to minimize the difference between the mean item effect between conditions. In such a case the assumption that the two sets of items have been sampled *independently* no longer holds.

The Problem and Proposed Solutions

The Standard F-test is Biased

Given that the analysis described above is relatively simple, one may wonder why such designs have led to so much debate. One of the reasons is that in a typical experiment with language materials and with latencies as the dependent variable, one almost never has a complete set of data. For example, in a lexical decision experiment, some items will be responded to incorrectly. Since reaction times are calculated using only the correct responses, there will be missing data. The usual approach is to compute for each subject the mean reaction time for each experimental condition and use these means in the ANOVA. Since such data appear to be identical to those of a standard design where one observes just a single value for the dependent variable, it is not surprising that researchers assumed that the same type of analysis could be used. The *F*-test statistic that would be used is that case is

$$F_1 = \frac{MS_A}{MS_{AS}} \qquad (2)$$

(note that MS_A and MS_{AS} have the same value as in the original analysis described in Table 1).

Such an approach would in fact have been correct, if each participant had been presented a different set of items (the random effect of items then becomes part of the within-cell error term). However, if each participant gets the same set of items, this is no longer true. The crucial difference is that if the same set of items is given to each subject, chance fluctuation in the mean effect of items will not average out when averaging over subjects. In that case, F_1 as defined above is severely biased as may be seen if we substitute the expected values for MS_A and MS_{AS} from Table 1 into Equation 2:

$$E(F_1) \approx \frac{E(MS_A)}{E(MS_{AS})} = \frac{\sigma^2_e + \sigma^2_{W(A)S} + q\sigma^2_{AS} + n\sigma^2_{W(A)} + nq\sigma^2_A}{\sigma^2_e + \sigma^2_{W(A)S} + q\sigma^2_{AS}} \quad . \qquad (3)$$

Clearly, even when $\sigma^2_A = 0$, F_1 may still be significant due to the $\sigma^2_{W(A)}$ term being larger than zero.

The Quasi F-ratio Cannot be Computed

Coleman (1964) and Clark (1973) argued that in order to test the null hypothesis of no differences between the treatment conditions, a *quasi F-ratio* or *F'-ratio* had to be used, that is, a statistic that has the form of an *F*-ratio but is not a true *F*-statistic (it is not based on independent sums-of-squares). For this design the *F'* has the following structure:

$$F' = \frac{MS_A + MS_{W(A)S}}{MS_{AS} + MS_{W(A)}} \quad . \qquad (4)$$

F' has an approximate *F*-distribution with degrees of freedom for the numerator and the denominator that may each be calculated by applying the following general formula:

$$df = (MS_1 + MS_2)^2 \,/\, \left(MS_1^2 / df_1 + MS_2^2 / df_2\right) \qquad (5)$$

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).

Unfortunately, in most situations *F'* cannot be computed due to missing data (error responses). Clark (1973) therefore advocated the use of a lower bound for *F'* termed *minF'*. *MinF'* was defined as

$$minF' = \frac{MS_A}{MS_{AS} + MS_{W(A)}} \qquad (6)$$

(note that *minF'* is always smaller than *F'* and equals *F'* if $MS_{W(A)S} = 0$).

The simplest way to obtain all of the quantities required for the calculation of *minF'* is to run *two* ANOVAs: one in which one collapses or averages over items (usually termed a *subject analysis*) and one in which one collapses over subjects (termed an *item*

analysis). In fact, it may be shown that the *minF'* statistic is equal to $(F_1F_2)/(F_1+F_2)$ where F_1 is the *F*-ratio for the treatment effect in the subject analysis and F_2 is the *F*-ratio for the treatment effect in the item analysis (see Clark, 1973).

Shortly after Clark's paper was published, there was a discussion in which some critics (Smith, 1976; Wike & Church, 1976) rejected the use of the *F'* (and *minF'*) as being an unduly conservative test. However, Monte Carlo simulations with *F'* as in Equation 4 have demonstrated that it is a good approximation to the normal *F* statistic and not unduly conservative, given that the variance components $\sigma^2_{w(A)}$ and σ^2_{AS}, expressing item and subject variability, are not too small (Davenport & Webster, 1973; Forster & Dickinson, 1976). In addition, Santa, Miller, and Shaw (1979) demonstrated that the *F'* is robust against violations of homogeneity and normality.

Confusion in the Interpretation of F_1 and F_2

In hindsight it appears unfortunate that the statistics F_1 and F_2 were introduced as a means to compute the value of *minF'*. Although researchers (especially in the language area) did take the fact that items are a random factor seriously and followed Clark's advice to calculate F_1 and F_2, confusion soon emerged about the exact nature of the problem and Clark's solution to it. As documented by Raaijmakers, Schrijnemakers, and Gremmen (1999), researchers started to treat F_1 as a test for generalization over subjects and F_2 as a test for generalization over items rather than as a intermediate result in the computation of *minF'*. Although initially the majority of papers that calculated F_1 and F_2 did at least also report *minF'*, the proportion of papers that did so steadily decreased in the years between 1975 and 1997 (see Figure 1 in Raaijmakers et al., 1999). By the end of that period almost no article published in the *Journal of Memory and Language* reported the *minF'* statistic.

One of the reasons for this may have been the fact that researchers noticed that sometimes F_1 and F_2 would be significant but *minF'* would not. Since a statistically significant result is still a major factor in getting one's results published, it is perhaps understandable that researchers began to view *minF'* as overly conservative (see Raaijmakers et al., 1999, for an example). However, it should be clear that this practice is without foundation since for the design given in Table 1, both F_1 and F_2 are biased; hence the fact that sometimes both of these statistics are "significant" but *minF'* is not, is to be expected. As mentioned previously, simulation results do not support the assertion that *minF'* is a too conservative test.

Overgeneralization to Other Designs

Presumably due to the misunderstanding of the relation of the statistics F_1 and F_2 to the "language-as-fixed-effect" problem, researchers also incorrectly generalized the solution to other designs than the standard one given in Table 1. Two designs in particular have been discussed in the literature: *matching* of items on relevant properties between conditions and *counterbalancing* of items across conditions. Both of these designs differ in important respects from the standard design in which items are randomly sampled within conditions. The consequences of matching of items were investigated by Wickens and Keppel (1983) who showed, using simulation results, that the bias in F_1 is greatly reduced when such matching is used. Wickens and Keppel (1983) discussed the case where individual items are matched using a blocked design. They showed (see also Raaijmakers et al., 1999) that if in such a design the blocking factor is ignored (i.e., the same analysis design is used as in Table 1), the best approach would be 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).

In order to understand why this is the case, it should be remembered that in the standard design it is assumed that the variability in the mean effect of items between conditions is directly related to the variability of items within conditions. That is, the term that occurs in the expected mean square for the treatment effect is the same as the corresponding term in the expected mean square for the Words within Treatments effect. However, if the items are matched between conditions, the variability (in replications of the experiment) in the mean item effect *between conditions* will be much reduced compared to the variability of the item effect *within conditions*. Hence, trying to compensate for the bias in F_1 using the term obtained from the Words within Treatments effect (as would be the case if one would compute *F'* or *minF'*), would be an overcompensation and hence too conservative.

In actual practice, however, researchers often do not use matching at the level of individual items. Rather, the two (or more) sets of items to be used in the two conditions are matched on a number of relevant variables by ensuring that the mean values for those variables (and sometimes the variances) are comparable in the two conditions. This might be termed a *set-matching* procedure since the two sets of items are matched, not the individual items. Following the reasoning above, it can be seen that in such a case too, the variability of the average item effect between conditions will be much smaller than that within each of the two conditions (if it were not, the matching would

TABLE 2
Expected Mean Squares for Repeated Measurements ANOVA With Counterbalanced Lists

Source of variation	df	Expected mean squares
G (groups) (= AxL 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	$\sigma_e^2 + np\sigma_L^2$
AxL (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.

not have been very successful). Hence, a set-matching procedure will also significantly reduce the bias in F_1 that would exist had the items been sampled completely randomly without any attempt to match the two sets of items.

Note that whether F_1 or *minF'* is to be preferred, depends on the extent to which matching has been successful. If the matching hardly reduces the variability in the difference in the mean item effect between the conditions, *minF'* will still be the preferred method of analysis. However, in most cases the matching will indeed result in a much reduced variability, and hence in most cases, F_1 should be the preferred analysis.

Matching of items (at least at the set level) is a common procedure that is used when the experimental effect to be tested involves some property of the words (such as natural language frequency or nouns versus verbs, etc.). In many experiments, however, it is possible to use the same items in both experimental conditions. For example, in a primed lexical decision experiment the same target words might be paired with an associatively related prime as well as with an unrelated prime. If that is the case, a counterbalanced design might be used in which two lists of items are constructed and 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. Such counterbalanced designs were discussed in detail in Pollatsek and Well (1995). Raaijmakers et al. (1999) showed that in such a design the treatment effect can be tested directly without the need to perform both a subjects and an item analysis, contrary to the suggestion of Pollatsek and Well (1995). Rather, for each subject one simply computes the mean value for each separate list. Note that in this design the random effect of words is taken into account through the assumption that in replications of the experiment, different lists might be used; hence the Lists factor is treated as a random effect.

The ANOVA model for this design is as follows:

$$X_{ijm(t)} = \mu + \theta_t + \pi_{m(t)} + \alpha_i + \beta_j + \alpha\beta'_{ij} + \varepsilon_{ijm(t)}, \quad (7)$$

where μ = overall mean, θ_t = effect of group t (= the between component of the Treatment x List interaction), $\pi_{m(t)}$ = effect of subject m (nested within group t), α_i = effect of the experimental treatment i, β_j = effect of list j, $\alpha\beta'_{ij}$ = the within component of the Treatment x List interaction, and $\varepsilon_{ijm(t)}$ = experimental error (a residual term equivalent to the interaction between Treatment, List, and Subjects plus "real" error). Due to the nature of this design (each group receives only p of the pxp combinations of Treatment and List), the interaction between Treatment and List is divided into two components, one between-subjects and one within-subjects. To see this, note that if one makes a table listing all pxp combinations of Treatment and List (which together comprise the interaction of Treatment and List), p of the cells correspond to Group 1, p to Group 2, etc. Thus, part of the differences that make up the interaction is equivalent to the differences between the groups, the main effect of Groups. Since Groups is obviously a between-subjects factor, this part of the interaction is conventionally termed the between-subjects part of the interaction, and the remaining part is called the within-subjects part of the interaction.

In the ANOVA model, it is assumed that Subjects within Groups as well as Lists are random factors (Groups might also be said to be random since it corresponds to an interaction between a fixed and a random effect). Table 2 gives the expected mean squares for this design under these assumptions. Note that the

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 conditions, there will be no contribution to the variability between conditions due to lists and hence due to the main effect of items. Note that there will be an effect due to the interaction between items and conditions (some items might show a larger treatment effect than other items). This is reflected in the Lists x Treatments interaction component $n\sigma^2_{AL}$ in the expected mean squares for the treatment effect, as shown in Table 2. However, since that term also occurs in the expected mean squares for the A x L (within) effect, the treatment effect can be directly tested using the A x L (within) effect as the error term.[2] Raaijmakers et al. (1999) give all the necessary details for the computation of the required mean squares listed in Table 2.

Criticisms and Evaluation

What if the Effect Occurs Only for a Subset of the Items?

Despite the fact that the results summarized above have been known for some time and have resulted in clear recommendations for how to proceed in any given case, there still seems to exist a resistance among language researchers to change the existing practice of always performing both subjects and item analyses without computing *minF'*, even though there is no statistical rationale for such a procedure. Over the past couple of years, I have seen many instances, in some cases because researchers contacted me with specific questions and in other cases because I received copies of the comments that reviewers made on manuscripts that did not follow the standard (but incorrect) approach. Below I will give a few examples but first I would like to stress that the major problem still seems to be an incorrect or at least, imprecise, understanding of why it matters that items are a random effect and why sometimes item analyses should be performed. Some researchers seem to believe that item analyses are always necessary, irrespective of whether the *F*-test is biased or not, simply because there might be variability between the items and one has to perform an item analysis to determine whether the effect holds for all items rather than for a subset of the items.

For example, one reviewer noted:[3]

> My only substantive concern is about the item analyses. Perhaps the authors were making a little joke when they said in their letter that Reviewer A "noted" that Raaijmakers et al. have "shown" that you don't need item analyses in certain situations. I do not believe that Raaijmakers et al. have shown that, and literally (not virtually) every psycholinguist I know also does not believe that. If a small minority of items produces an effect, even with counterbalancing, you can get a significant result by running more subjects, but you can't get a significant effect in the item analysis that way. For this reason, item analyses are still required by JML and other psycholinguistic outlets.

Another reviewer voiced similar concerns:

> The claim that counterbalanced designs don't need item analyses… is absolutely untrue. Such a design removes one problem that normally requires an item analysis (inadequate matching of materials across conditions), but it does not eliminate the issue of generality. Consider the possibility that, say, 25% of the items show a priming effect, and the remainder show random effects. This could give you a very strong effect in the subject analysis, but a non-significant effect in the item analysis. A counterbalanced design doesn't help here. And if one is tempted to say that an effect in 25% of the items is still an effect, consider the mirror image case: an effect in only 25% of the subjects, and random effects for the remainder.

These quotes demonstrate a concern that one frequently encounters, although one rarely sees it in print. The idea seems to be that a treatment effect should be present for each and every one of the subjects and items, otherwise it is not a real effect, and that an item analysis shows whether or not this is the case. However, it is not difficult to show that such an idea is incorrect. First, it should be noted that if one tests for a treatment effect, one tests whether the population means for the two conditions are different. It should be clear that even if only 25% of the items (or subjects for that matter) show the effect, the population means will still be different. For example, if 25% of the items have an effect of 40 ms and the remainder none, the difference between the population means for the two conditions will be 10 ms, and any valid test should give a significant result given enough power. The standard *F*-

[2] It is generally advisable to first test whether there is a significant A x L (within) effect. If this test is not significant by a conservative criterion (say $\alpha = .25$), then the mean square for the A x L (within) effect may be pooled with the error (residual) mean square, giving a much more powerful test for the treatment effect (since the error term will be based on a large number of degrees of freedom). Note that if $p = 2$, the A x L (within) effect does not exist. In that case the treatment effect is always tested against the residual error mean square.

[3] Since I do not intend to criticize specific individuals, the identities of the persons who made specific comments is kept confidential.

TABLE 3
Example Data for Design Using Counterbalanced Lists

		Condition 1				Condition 2			
Group	Subj	item 1	item 2	item 3	item 4	item 5	item 6	item 7	item 8
1	1	505	491	498	566	503	497	493	500
	2	495	495	494	569	505	503	496	500
	3	497	496	501	549	510	498	492	498
	4	492	489	506	559	500	504	498	511
		item 5	item 6	item 7	item 8	item 1	item 2	item 3	item 4
2	5	557	558	559	512	491	494	504	503
	6	558	558	557	497	496	497	503	502
	7	561	567	570	508	487	504	493	495
	8	558	560	564	492	507	502	494	499

TABLE 4
Data From Table 3 Collapsed Over Items

		Condition 1	Condition 2
Group	Subj	List 1	List 2
1	1	515.00	498.25
	2	513.25	501.00
	3	510.75	499.50
	4	511.50	503.25
		List 2	List 1
2	5	546.50	498.00
	6	542.50	499.50
	7	551.50	494.75
	8	543.50	500.50

TABLE 5
ANOVA Summary Table for Example Data of Table 4

Source of variation	SS	df	MS	F
G (groups) (= A x L)	964.9	1	964.88	542.07
S(G)	10.7	6	1.78	0.13
A	3,592.5	1	3,592.50	263.09
L (lists)	1,273.6	1	1,273.60	93.27
error	81.9	6	13.66	

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 ideas appear to be relatively common, at least within the language community. A (constructed) example may show more clearly what is happening in such a case. In this example, we have constructed a set of data for a counterbalanced design with two conditions and two lists of items (see Table 3). The data are constructed in such a way that Items 4, 5, 6, and 7 have an effect of 60 ms and the remainder of the items have no effect at all. Apart from this, no other effects were included in the generated dataset except a small amount of random noise. If one uses the approach advocated by Raaijmakers et al. (1999), the data are averaged within the lists, which results in Table 4.

Table 5 gives the ANOVA summary table corresponding to this analysis (see also Table 2). Just as was to be expected, the treatment effect is significant as well as the List main effect and the interaction between List and Condition (which in this design is equivalent to the main effect of Groups). Note that the List factor is significant due to the fact that the crucial items (4-7) are not distributed evenly across the lists. If an item analysis is carried out on these data (even though such an analysis lacks a clear statistical rationale in this case), an F_2 value of 7.44 (df = 1,6) is obtained (assuming that the group factor is included in the analysis). Despite the claims to the contrary, it is not at all evident how one would be able to determine from this statistic that the effect is limited to a subset of the items.

A second example using matched items (or if that is experimentally feasible, the same items) in the two conditions is shown in Table 6. In this example, only the items (or blocks) 1-4 show a treatment effect (of 60 ms), while the remaining items show no effect at all. The corresponding ANOVA summary table is given in Table 7 (see also Table 5 in Raaijmakers et al., 1999). Since there are no missing data, it is in this case possible to compute the quasi F-ratio using the equations given in Raaijmakers et al. (1999, Equation 8). For these data, F' = 5.62 (df = 1,7). Using either the statistics from Table 7 or by running separate subjects and item analyses, we obtain F_1 = 226.1 (df = 1,3) and F_2 = 5.75 (df = 1,7). This leads to $minF'$ = 5.61 (df = 1,7). Thus, even though the effect is only present for a subset of the items, it is significant at the .05 level, for F' as well as for $minF'$. Note that this analysis also shows that the extremely large interaction component Items x Treatments that was introduced in this constructed example does lead to a large bias in F_1 (see Equation 9 in Raaijmakers et al., 1999).

What such comments do show is that the interpretation of the item analyses has changed from a tool for the computation of $minF'$ to a tool for testing whether

TABLE 6
Example Data for Repeated Measurements ANOVA With Items Crossed With Treatment

condition	subj	item 1	item 2	item 3	item 4	item 5	item 6	item 7	item 8
1	1	505	491	498	506	503	497	493	500
	2	495	495	494	509	505	503	496	500
	3	497	496	501	489	510	498	492	498
	4	492	489	506	499	500	504	498	511
2	1	557	558	559	572	491	494	504	503
	2	558	558	557	557	496	497	503	502
	3	561	567	570	568	487	504	493	495
	4	558	560	564	552	507	502	494	499

TABLE 7
ANOVA Summary Table for Example Data of Table 6

Source of variation	df	Mean Square
A (Treatment)	1	14,370.02
B (Items)	7	2,163.57
S (Subjects)	3	1.35
A x B	7	2,497.12
A x S	3	63.56
B x S	21	33.50
A x B x S (residual)	21	29.09

the effect is specific to certain items rather than all items. Now the original (and correct) use of item analyses did of course have something to do with the generality of the observed result, but the generality that was meant there had to do with the generalization of the difference between treatment means to the population from which these items were sampled, not the generalization to each and every individual item.

To reiterate, the idea that the inherent variability between items always necessitates doing an analysis over items *irrespective of the actual nature of the design* appears to be virtually ineradicable. Many language researchers seem to be routinely applying statistical procedures without regard for the actual nature of the design that is used. Item analyses may be interesting in their own right but it should not be assumed that such analyses should always be used for testing the treatment main effect, irrespective of the properties of the design.

What if the Restrictions Used Exhaust the Population of Items?

Another objection is that in a particular experiment the items may have been selected subject to so many restrictions that the items that have been used virtually exhaust the population. Therefore, the Item factor cannot really be considered random since that would imply a virtually infinite population of items that could have been used (hence, Items should be treated as a fixed effect). This might indeed be true given the restrictions used, but the real question is whether the researcher would indeed refuse to consider a "replication" of the experiment in which a different set of restrictions was used (e.g., a different range of frequencies) as a valid replication. And what about a replication of the experiment in a different language? These are real problems to which there is no easy answer but that should be considered before jumping to the conclusion that Items should be treated as fixed. A variant of this is when a researcher does admit that there may be other samples of items that could have been used but believes that the sample actually used in the experiment represents a sizeable fraction of all possible items (e.g., when one used 20 items in the experiment and one estimates that the population contains 50 potentially suitable items in total).

It is indeed true that such cases violate the assumptions underlying the random effects model. One solution would be to use appropriate correction factors in the formulas for the expected mean squares (known as the finite population correction). However, in my opinion one should be cautious in using such an approach. It is often the case that a researcher uses a particular set of restrictions but does believe that the conclusions should hold more widely. An analogy that illustrates the problem is the following. Suppose that a researcher does an experiment on spatial attention in air traffic control. In order to get participants, he contacts a nearby airport and succeeds in obtaining the cooperation of all 20 air traffic controllers from that airport. Should he then treat the subjects factor as fixed rather than random? Probably not. It is more likely that he will assume that his group of air traffic controllers are sufficiently representative for air traffic controllers in general (at least with respect to the experimental treatment effects). Note that this is an example of the general rule

that in order to answer questions regarding the random or fixed character of experimental effects, one has to consider the set of valid replications of the experiment. If there is no a priori reason to assume that items (or subjects) could not have been selected in a different way, using different restrictions, it is probably best (i.e., the most conservative) to treat the Item factor as random rather than fixed.

How do You Know That the Matching of Items was Successful?

A third objection is that even though careful matching may have been used, one can never be sure that the items or the lists have been matched on all relevant dimensions. In particular, it may be the case that the matched items differ with respect to the treatment effect even though they may be comparable overall. There is no easy solution in this case. On the one hand, it might be better to assume complete matching (especially since the main effect of items will usually be larger than the interaction between items and treatment); on the other hand, it could in principle also be defended to assume that the matching does not significantly reduce the variation due to items (even though one would be tempted to question why the researcher took the trouble to search for matching pairs if that does not really accomplish what it was supposed to do).

The problem is that if the matching is only partially successful, the *minF'* test will be too conservative but the F_1 test will be too liberal. There might be a solution to this problem, even though it may involve some extra work by the researcher. Let us again assume that there are two experimental conditions, and that we are able to select two samples of 20 words, one for each condition (i.e., words are nested within conditions) where the two samples are matched (in terms of their mean values) on a number of relevant dimensions. If that is possible, it might also be possible to select four lists of 10 items each, two for each condition, that are also matched on those dimensions. Let us denote the two lists for Condition 1 as List 1 and List 2 and the two lists for Condition 2 as List 3 and List 4.

In the proposed analysis we will use the mean score for List 1 as the dependent variable. If words are assumed a random effect, then this new list factor will also be a random effect. Note that even though partial matching may have been used, *the variability between lists within a particular experimental condition will still be comparable to the variability between conditions*. Hence, in such a design the assumptions for the standard design (see Table 1) are met, provided that one substitutes Lists for Items. We could therefore do an ANOVA with subjects and lists as random effects using the scheme described in Table 1. In all respects, the design is exactly the same as in the case considered by Clark (1973) except that "words" has been replaced by "lists."

One objection might be that we now have only two "items" per cell; hence there will be a loss in the degrees of freedom and hence a loss in power. However, it should be realized that the variability between the lists, as defined here, will be substantially smaller than the variability between words, and this will at least partly compensate for the decreased number of "items." In addition, the error variability will of course be reduced since we are using averages. In addition, such a procedure eliminates the problem of missing data and hence we can now compute *F'* directly rather than its lower bound *minF'* (see Equation 1).

More generally, the number of items within each condition may be divided into as many lists as are feasible, and are still sufficiently large to ensure that no subject has all missing scores for any of the lists (i.e., there are no missing data in the analysis in which the data have been reduced to list means per subject).

Although we have not carried out a full simulation study to study the properties of this approach (and in particular its power), a few numerical examples show that the approach at least deserves further analysis. For example, using artificial data in a design with 2 treatment conditions, 10 subjects and 2 lists of 5 items each in each condition, in which 80% of the item variance was matched between lists, the analysis based on the average scores per list showed a treatment effect that was just significant, $F'(1,3) = 9.17$, $p < .056$, while the analysis in which the items within a list were not averaged, did not show a significant effect, $F'(1,20) = 2.86$, $p < .106$ and $minF'(1,20) = 2.80$, $p < .110$. However, other examples in which the proportion of matched variance was lower showed a reversed pattern (i.e., the *F'* for the analysis based on "Lists" was less significant than the *minF'* for the analysis based on the individual items). At this moment it is not clear exactly when the one analysis is to be preferred over the other.

Thus we are left with the somewhat unsatisfactory conclusion that the proposed analysis may provide a better way to test treatment effects than the standard *minF'*-test, especially when the item variance is large and a substantial portion of the item variance is controlled through matching, but the precise conditions that have to be fulfilled are not yet clear.

Conclusion

I conclude by summarizing the main recommendations. First, in those cases where items cannot be counterbalanced across conditions and matching of items

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 the item scores. Finally, if the experiment allows counterbalancing of items across conditions the correct procedure is the one described in Raaijmakers et al. (1999) since that procedure allows a direct test of the main hypothesis even if items are assumed to be a random effect.

I would like to thank Peter Dixon, Stephen Lupker, Jos van Berkum, Diane Pecher, and Eric-Jan Wagenmakers for helpful comments on an earlier version of this article. Correspondence regarding this article should be addressed to Jeroen G.W. Raaijmakers, Department of Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands (E-mail: J.G.W.Raaijmakers@uva.nl).

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.

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.

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.

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with "The language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language*, *41*, 416-426.

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.

Sommaire

L'analyse convenable des expériences à l'aide d'un matériel langagier est source de controverse et de débats entre les chercheurs. Le problème est soulevé de façon tout à fait éloquente dans une conception d'expérience où les mots sont enchâssés dans les conditions de traitement, par exemple en comparant des noms et des verbes ou des occurrences élevées par opposition à faibles. Dans une telle conception la variabilité de l'échantillonnage des items contribue au carré moyen du facteur de traitement. Il s'ensuit que le test F standard (obtenu en établissant la moyenne des items) sera biaisé. Clark (1974) préconisait le calcul de la donnée statistique minF' pour éliminer ce biais. Même si cette solution a fait l'objet de nombreuses discussions, il existe encore de la confusion entourant son utilisation adéquate et particulièrement en ce qui touche les conditions dans lesquelles ces données statistiques sont adéquates (en terme de biais et de puissance).

Un certain nombre d'objections contre l'utilisation de minF' sont abordées. Nous montrons qu'une telle procédure ne sera pas adéquate si la variabilité du score moyen des items entre les conditions est (beaucoup) moindre que ce à quoi on pourrait s'attendre selon la variabilité de la condition intrinsèque. Une telle situation survient si les items ont été appariés entre les conditions d'un certain nombre de propriétés, soit au niveau individuel ou au niveau de la moyenne de chacune des conditions. Dans de tels cas les données statistiques de minF' seront trop prudentes. Dans le même ordre d'idée, si les items ont été contrebalancés entre les conditions, il faudrait tenir compte de l'aspect de cet équilibrage dans l'analyse statistique. Raaijmakers, Schrijnemakers et Gremmen (1999) décrivent comment de telles données devraient être analysées en tenant compte du fait que les items et les sujets sont aléatoires dans le compte rendu.

Nous décrivons un certain nombre d'objections qui ont été soulevés par rapport à ces recommandations. La plus importante étant la croyance (erronée) que l'approche consistant à contrebalancer les données préconisée par Raaijmakers et al. (1999) ne pallie pas au

besoin d'effectuer des analyses d'items distinctes étant donné qu'il pourrait se trouver un item par interaction de traitement qui soit tel que l'effet ne se produit que pour un sous-ensemble d'items. Il semblerait qu'on croit que si c'est le cas, l'effet du traitement ne devrait pas être significatif. Cependant, cette supposition est fausse étant donné que même lorsqu'un effet ne se produit que pour un sous-ensemble d'items du groupe, il restera tout de même un effet de traitement principal dans le groupe et tout test suffisamment sensible devrait montrer un effet principal significatif pour le traitement. Évidemment, l'item de l'effet d'interaction par traitement (s'il existe) devrait aussi être significatif, mais que ce soit ou non le cas, il n'indique à peu près rien sur la différence entre les moyens de traitement du groupe. Nous illustrons ces problèmes à l'aide de deux ensembles de données simulées. Nous abordons aussi brièvement ce que serait la meilleure pratique si les items étaient choisis en fonction d'un nombre de restrictions si grand que les items utilisés dans l'expérience épuisent à toute fin pratique le groupe. En dernier lieu, nous décrivons une solution possible pour le problème que l'on ne peut jamais être certain que l'appariement qui a été utilisé est tel que les items ont en effet été appariés en fonction de toutes les dimensions pertinentes.