*Behavior Research Methods*
*2007, 39 (4), 735-747*

# Analysis of variance for repeated measures designs with word materials as a nested random or fixed factor

**Toni Rietveld and Roeland van Hout**
*Radboud University Nijmegen, Nijmegen, The Netherlands*

This article is about analysis of data obtained in repeated measures designs in psycholinguistics and related disciplines with items (words) nested within treatment (= type of words). Statistics tested in a series of computer simulations are: $F_1$, $F_2$, $F_1$ & $F_2$, $F'$, min $F'$, plus two decision procedures, the one suggested by Forster and Dickinson (1976) and one suggested by the authors of this article. The most common test statistic, $F_1$ & $F_2$, turns out to be wrong, but all alternative statistics suggested in the literature have problems too. The two decision procedures perform much better, especially the new one, because it systematically takes into account the subject by treatment interaction and the degree of word variability.

Repeated measures designs (also called *within-subjects designs*) are frequently used in disciplines like psycholinguistics, phonetics and speech and language pathology. They are potentially very powerful, although only a relatively small number of subjects (participants) is required, as all subjects have to react to all items. We will use the term subject instead of participant to avoid any misunderstanding about the more technical terminology of within- and between-subjects effects or factors. In this contribution we address questions that are relevant in the context of a repeated measures design that is common in psycholinguistic research—that is, a design with word materials nested within type of words (nested within treatment). Many aspects of this design have been the topic of debate and of continuing research. Aspects discussed include the following: (1) the question as to whether items ("words") used in current psycholinguistic experiments should be considered random or fixed effects and the associated problem of finding statistics that yield a fair balance between power and the chance of making Type I errors (Clark, 1973; Coleman, 1964); (2) the choice of design and test statistics with sampled word materials (Raaijmakers, Schrijnemakers, & Gremmen, 1999; Wickens & Keppel, 1983); (3) possible advantages of alternative approaches, such as multilevel analysis (Quené & van den Bergh, 2004). Of course, analysis of variance is not the only approach to the analysis of repeated measures data. Multilevel modeling (also known as *hierarchical linear models*) is a frequently used alternative, which is particularly useful when missing data are present, as is often the case in experiments with reaction times as outcome variable (Quené & van den Bergh, 2004). Yet, analysis of variance is still the default analysis instrument, and this is the case not only because researchers are used to it. One of the potential problems associated with the multilevel approach occurs in the analysis of small samples (see Gueorguieva & Krystal, 2004). Since the number of subjects in repeated measures designs is often relatively small, we expect analysis of variance to be the default approach to repeated measures for the years to come. Moreover, it has still to be sorted out how multilevel analysis can deal with a within-subjects factor that is neither random nor fixed.

Clark, in his influential 1973 article, started a whole series of contributions by statisticians and psychologists examining the validity of test statistics when language materials are used. Following Clark's advice, researchers now consider words as a random factor, also in an attempt to ensure generalizability of their results. To test the significance of the factor treatment, they resort to the combined use of $F_1$ (in which words are seen as a fixed and subjects as a random factor) and $F_2$ (in which the assignment of random and fixed is the other way around). This use of $F_1$ and $F_2$ has become standard practice in psycholinguistic research, even though it is not in accordance with Clark's recommendation to use $F'$ or, alternatively, min $F'$ when language materials constitute a random factor.

We will deal with the question of why the combined use of $F_1$ and $F_2$ has become a standard in psycholinguistic research, and in relation to this we discuss two important and influential post-Clark articles: Forster and Dickinson (1976) and Wickens and Keppel (1983). Forster and Dickinson showed the enormous impact of a subject-by-type (type = treatment) interaction in combination with word variability on different test statistics. Wickens and Keppel made clear that the degree of systematic sampling needs to be related to the choice of a test statistic. In a series of

**T. Rietveld, a.rietveld@let.ru.nl**

**Table 1A**
**The Classic Repeated Measures Design, With *r* Subjects (S) and With Words (W) Nested Within Type (T)**

| | $T_1$ | | | $T_2$ | | | $T_p$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $W_1$ | …… | $W_q$ | $W_{q+1}$ | …… | $W_{2q}$ | $W_{(p-1)q+1}$ | …… | $W_{pq}$ |
| $S_1$ | | | | | | | | | |
| . | | | | | | | | | |
| . | | | | | | | | | |
| . | | | | | | | | | |
| . | | | | | | | | | |
| $S_r$ | | | | | | | | | |

simulation experiments, we will assess, under different statistical conditions, the behavior of $F_1$, $F_2$, $F'$, min $F'$, the conventional decision procedure of the $F_1$ *and* $F_2$ ($F_1$ & $F_2$, in our terminology) combination and the decision procedure suggested by Forster and Dickinson. We will also put our own decision procedure to the test, which explicitly takes word variability and the subject-by-type interaction into account. Particular attention will be paid to the question of whether the assumption that "randomness" and "fixedness" is a dichotomy is valid, and we will also discuss the statistical and methodological consequences of the subject-by-type interaction. Furthermore, we will discuss the effects of missing data on the statistical outcomes, because a major reason for adhering to the $F_1$ & $F_2$ decision procedure seems to be the easy way this rule allows the researcher to disregard missing data.[1]

## CURRENT PRACTICE: $F_1$, $F_2$, AND MIN $F'$

What is the current practice of analyzing and reporting experiments with repeated measures on subjects (S) who have to judge or react to words (W) that are nested within a type (T)? The design of such an experiment is given in Table 1A, with $r$ subjects, $p$ types, and $q$ items per type. The sources of variation and the expected values of their mean squares are given in Table 1B.

If both words and subjects are regarded as random factors—which most researchers prefer—no error term can be found to test the factor type (T). One remedy, the $F'$ ratio, can be calculated by summing mean squares both in numerator and denominator in such a way that the expected value of the former differs from the latter by just one term (= the variance associated with the factor T): $F' = (MS_T + MS_{S\times W(T)})/(MS_{S\times T} + MS_{W(T)})$. To our knowledge, however, no statistical software currently available offers the possibility of calculating an $F'$ directly. This seems to be one of the reasons why most researchers tend to report other statistics: $F_1$ and $F_2$. Clark distinguished two $F$ values in order to test the effect of the factor type: $F_1$, in which words is considered a fixed factor (and subjects a random factor), and $F_2$, with words as a random factor (and subjects as a fixed factor). Since the expected values of the mean squares vary as a function of the character of the factors (random/fixed), both $F$ values are obtained on the basis of different error terms: $F_1 = MS_T/MS_{S\times T}$, and $F_2 = MS_T/MS_{W(T)}$. Most statistical software packages do not provide facilities to carry out analyses of variance on repeated measures data with different fixed/random labels for the factors. Moreover, and more problematically, the calculation of $MS_{S\times T}$ and $MS_{W(T)}$ requires a *complete* data set, which is seldom available. Fortunately, Clark showed that $F_1$ and $F_2$ can also be obtained by simply pooling over words (items) or subjects, respectively. That is why $F_1$ is also called $F_s$ ($F_{subjects}$) and $F_2$ is called $F_i$ ($F_{items}$).[2]

Pooling over items results in a randomized block design, with one within-subjects factor, T. The $F$ ratio for factor T is the same as $F_1$. Pooling over subjects results—in our example—in a one-way analysis of the mean values of the words, obtained by averaging over the subjects. A one-way analysis (with type as the factor) applied to these data yields the same $F$ ratio as $F_2$. An advantage of the use of $F_1$ and $F_2$ is that it enables the researcher to calculate min $F'$ quite easily: $MS_T/(MS_{S\times T} + MS_{W(T)})$. This statistic, the lower-bound value of $F'$, is recommended by Clark (1973) and Coleman (1979) in order to avoid Type I errors. The easy-to-calculate version of min $F'$ is $(F_1 \times F_2)/(F_1 + F_2)$.

In most articles on research in which repeated measures designs are used with word materials nested within the factor type (= treatment), only $F_1$ and $F_2$ are reported. As Wickens and Keppel (1983) and Raaijmakers et al. (1999) pointed out, both statistics are positively biased in many situations and consequently result in high Type I error rates. The ratios of mean squares used for these statistics are only unbiased if we can be sure that for $F_1$, $\sigma^2_{W(T)} = 0$, and for $F_2$, $\sigma^2_{S\times T} = 0$ (see Table 1B), both of which obviously do not happen very often. Wickens and Keppel as well as Raaijmakers et al. therefore recommend the use of $F'$ in those cases in which the language materials can be considered as being randomly sampled from a large population. Although $F'$ is an approximation of the genuine $F$

**Table 1B**
**Expected Values of Mean Squares for the Design Given in Table 1A**

| Source | *df* | E(*MS*) |
|---|---|---|
| S | $r-1$ | $\sigma^2_\varepsilon + [(Q-q)/Q]\,\sigma^2_{S\times W(T)} + pq\sigma^2_S$ |
| T | $p-1$ | $\sigma^2_\varepsilon + [(Q-q)/Q]\,\sigma^2_{S\times W(T)} + q\sigma^2_{S\times T} + [(Q-q)/Q]r\sigma^2_{W(T)} + qr\sigma^2_T$ |
| W(T) | $p(q-1)$ | $\sigma^2_\varepsilon + [(Q-q)/Q]\,\sigma^2_{S\times W(T)} + r\sigma^2_{W(T)}$ |
| S×T | $(p-1)(r-1)$ | $\sigma^2_\varepsilon + [(Q-q)/Q]\,\sigma^2_{S\times W(T)} + q\sigma^2_{S\times T}$ |
| S×W(T) | $p(q-1)(r-1)$ | $\sigma^2_\varepsilon + \sigma^2_{S\times W(T)}$ |

Note—$(Q-q)/Q$ is 0 when the factor word [= W(T)] is fixed, and 1 when this factor is random. $Q$, number of levels of words in the population; $q$, the associated number in the sample; S, subjects; T, type.

ratio, Wickens and Keppel made clear that tests based on $F'$ are correct if the word materials—and subjects—are sampled in a random way. Moreover, Maxwell and Bray (1986) showed the robustness of $F'$ to the common risk in repeated measures designs: the violation of sphericity. Thus basing ourselves on the literature, nothing prevents us from applying $F'$ when language materials do constitute a random factor. Nevertheless, the convention of reporting both $F_1$ and $F_2$ prevails in most journals.

The present situation has been thoroughly investigated by Raaijmakers et al. (1999), who rightly warned of "the $F_1 \times F_2$ fallacy." To illustrate how widespread this fallacy is, they screened five volumes (1993–1997) of the *Journal of Memory and Language.* A total of 124 relevant articles were found, of which 120 reported $F_1$ and $F_2$ only, and of which 4 reported min $F'$ as well as $F_1$ and $F_2$. By investigating a longer time span (i.e., from 1974 onward), they convincingly pointed out a gradual change from a situation in which min $F'$ was the standard toward the present situation in which the $F_1$ & $F_2$ criterion is used exclusively. Raaijmakers et al. argued that in the case in which word variability is random, min $F'$ is the proper statistic. The $F_1$ & $F_2$ criterion is positively biased (resulting in a Type I error that is too high) when words are random, as was shown by Forster and Dickinson (1976) in a series of Monte Carlo simulations. Raaijmakers et al. claimed that in practice the traditional $F_1$ is the correct test statistic when word variability in experiments is kept under control, which especially applies in the case of item matching and counterbalancing. Their advice does not cover, however, the common situation in which word variability floats somewhere between real randomness and strict experimental control (fixed). Sampling three-syllabic words, for instance, gives a constrained set of possible words, and the reseracher has to take the test materials from the available language-specific (sub) set. The degree of experimental control is determined by the size and properties of the set of words available. Sizes and properties of lexical (sub)sets are defined by the research questions, and when these questions, for instance, are delimited to homophones or homographs between language pairs, the resulting subset may be extremely small.

## CHARACTERISTICS OF WORD MATERIALS

### The Role of the Variance Components $\sigma^2_{S\times T}$ and $\sigma^2_{W(T)}$

Forster and Dickinson (1976) carried out a number of Monte Carlo simulations on the classic repeated measures design in psycholinguistic research, as discussed in the previous section. Two variance components were systematically varied, which are characteristic of the sources of variation that figure as error terms in the $F_1$ and $F_2$ analysis when the full design is used: $F_1 = MS_T/MS_{S\times T}$, and $F_2 = MS_T/MS_{W(T)}$. Forster and Dickinson did not deal with these two components from an interpretive, but merely from a purely statistical point of view. An important initial conclusion they drew is that the reduction of word variability ($\sigma^2_{W(T)}$) clearly has a favorable effect on the statistical properties of $F_1$. When word variability is zero, $F_1$ is the test statistic required, for zero variability exactly defines a fixed word effect. If word variability is present, $F_1$ becomes too liberal.

A striking effect of a zero interaction of subject-by-type (the component $S\times T$) is the proper statistical behavior of $F_2$ in terms of Type I and Type II errors, independent of the presence or absence of word variability. These good properties disappear when the subject-by-type interaction effect increases, however. The power of $F_2$ depends on the number of items, of course, which means that a design with only a few items (words) will have insufficient power. Forster and Dickinson calculated Monte Carlo estimates for $F_1$, $F_2$, $F_1$ & $F_2$, min $F'$, and $F'$, systematically manipulating the variance components $\sigma^2_{S\times W(T)}$ and $\sigma^2_{W(T)}$. We summarize their results in Table 2 according to the presence or absence of these two variance components.

Table 2 illustrates that $F_1$ & $F_2$ only performs appropriately in the "either–or" situation—in other words, the situation in which either $\sigma^2_{W(T)}$ or $\sigma^2_{S\times T}$ is present. $F'$ performs well under all circumstances, the exception being when both variance components are zero. The same goes for min $F'$, but this statistic tends to be too conservative in other cases as well. The main conclusions to be derived from the results in Table 2 are (1) $F_2$ performs well as long as there is no interaction S×T; (2) since the presence of $\sigma^2_{W(T)}$ reflects item variability, and consequently the fixed or random character of the associated factor, $F_2$ appears to be insensitive to the question of whether the factor word is random or fixed; (3) the statistics $F_1$ & $F_2$, $F'$, and min $F'$ are too conservative if both variance components involved are absent. However, Forster and Dickinson (1976, p. 138) tell us not to worry too much, because in their simulations the nominal level of $\alpha$ was only affected moderately (from 5% to about 2.6% for $F'$, to about 1% for min $F'$, to about 2.2% for $F_1$ & $F_2$).

The variance components under discussion here, $\sigma^2_{W(T)}$ and $\sigma^2_{S\times T}$, are determinant factors for the behavior of the $F$ statistics. It is surprising, therefore, that the meaning of these components has been largely neglected, both in the statistical literature and in the interpretation of the results obtained in psycholinguistic experiments with repeated measures. These two components will be central in our statistical simulations.

### Fixed or Random Effects: Gradual Matching

Two standard strategies are in use to handle sets of stimuli in a (quasi)experimental design: (1) random sampling

**Table 2**
**Summary of the Forster and Dickinson (1976) Results As a Function of the Presence (= 1) or Absence (= 0) of the Variance Components $\sigma^2_{W(T)}$ and $\sigma^2_{S\times T}$**

| $\sigma^2_{W(T)}$ | $\sigma^2_{S\times T}$ | $F_1$ | $F_2$ | $F_1$ & $F_2$ | $F'$ | min $F'$ |
|---|---|---|---|---|---|---|
| 0 | 0 | OK | OK | − | − | − |
| 0 | 1 | OK | + | OK | OK(–) | OK(−) |
| 1 | 0 | + | OK | OK | OK | OK |
| 1 | 1 | + | + | + | OK | OK |

Note—The error rates of the $F$ ratios for a type/treatment effect were calculated under the condition of no effect. OK, error rate of about 5%; −, error rate is too low (<3%); OK(−), tendency to be too low (between 3% and 4%); +, error rate is too high (>7%).

and (2) matching on item or word level. We will briefly review both strategies, assuming that the topic of research is the difference in how two word-class categories, nouns and verbs, are processed by human subjects.

1. *Random sampling*. Random samples of items are drawn, for instance 10 verbs and 10 nouns. The items drawn within the two word class categories are nested; they only occur within the word class category they belong to. A disadvantage of this strategy is the necessity of drawing fairly large samples of items, especially when the variance within the populations (the word-class categories) is large.

2. *Item matching (= blocking)*. For every noun item, a comparable verb item is selected. The word *work*, for instance, can be both noun and verb. The question is whether phonological similarity is a sufficient condition. It does not guarantee, for instance, that the frequencies of the noun and the verb items are similar. Moreover, many verbs and nouns are excluded from selection. In practice, it is rarely possible to get a completely satisfactory match per item. Anyway, if item matching can be carried out in a psycholinguistic experiment, the analysis of variance is self-evident in the sense that items can be included as a *blocking* factor (every item fully crossed with type—e.g., word class). The item levels ought to be treated as a random factor. The corresponding expected mean squares when items are random can be found in Wickens and Keppel (1983, p. 306; see also Raaijmakers et al., 1999).

Both the strategy of random sampling and the strategy of item matching are hardly ever used in psycholinguistic practice. There is a strong preference to use a mixed strategy of matching on the level of sets of items in combination with a deliberate item choice. Items in the noun category are compared as a set to a set of verb items. The researcher wants to compare the two sets with regard to their frequency distribution, their syllable structure properties, their phonological form, and so on. The aim is to neutralize potentially confounding factors at the level of the items. The consequence of this is that the items selected are not randomly chosen but are selected deliberately, in such a way that relevant characteristics at the set level are similar, which directly raises the question of whether items/words should be interpreted as a random factor. In any case, matching at set levels has become standard in psycholinguistic research, and therefore it deserves a separate label: "set matching."

3. *Set matching*. Two or more sets of items are deliberately and carefully selected in such a way that potentially confounding factors are neutralized. The neutralization effect is established by obtaining comparable values on relevant variables at the set level. Wickens and Keppel (1983, p. 306) summarized this approach as a "mixture of sensible selection, counterbalancing, informal blocking, and the elimination of extreme material." Nevertheless, the factor word is nested, and it is unclear whether it should be given the status of a fixed or random factor.

### Missing Data

The $F_1$ & $F_2$ procedure seems to be an elegant way to disregard missing values. However, avoiding the problem of handling missing values does not mean that no choice has been made. Mean values are imputed implicitly, on the basis of pooling over items in $F_1$ and pooling over subjects in $F_2$. There are a number of ways in which missing data can be dealt with (cf. Allison, 2002; Gornbein, Lazaro, & Little, 1992; Little & Rubin, 1987; Rietveld & van Hout, 2005). The procedure described in Winer, Brown, and Michels (1991, pp. 487–490) has been the most frequently used method reported in psycholinguistic literature. Below, we propose the use of an alternative procedure ("hot deck").

## DESIGN-DEPENDENT STATISTICS AND MULTISTAGE PROCEDURES

Raaijmakers et al. (1999) concluded that the traditional $F_1$ leads to appropriate results in many cases, with the exception of designs in which items are fully nested within treatments. In such circumstances, (min) $F'$ is the statistic required. In fact, they claimed a larger area of application of $F_1$ than is commonly assumed in psycholinguistic research. They argued that the balancing of language materials leads to a model equation with subjects as random factor and materials as fixed factor. Wickens and Keppel (1983) came to the same conclusion on the basis of a series of simulations, which showed that balancing in general had a positive effect on the statistical properties of $F_1$. Their final conclusion was that the choice of the test statistic ought to depend on the underlying model one wants to assume. Why did the conclusion of Wickens and Keppel hardly have any impact on the statistical practice in psycholinguistic research? And why can the same be said for Raaijmakers et al.'s recommendations? One reason is that both sets of recommendations are difficult to handle in practice. They do not provide an answer to the question of how one should assess whether the properties of the word materials satisfy the conditions of a fixed factor. In such a situation, a researcher will take the safe track of double checking via $F_1$ and $F_2$. Paradoxically, this track is not safe at all.

If one accepts that no panacea is available, an attractive alternative is a multistage procedure. A well-known, but not frequently used, multistage procedure is the one given by Forster and Dickinson (1976). Since both $F'$ and min $F'$ are conservative tests, they suggested the following procedure for avoiding unacceptably high Type I errors.

### Multistage Procedure of Forster and Dickinson (1976; Hereafter, F&D)

(A) Test min $F'$. If it is significant, reject $H_0$. If it is not significant, proceed to Step B.

(B) Test the effects of S×T and W(T). These effects are tested by dividing their *MS*s by $MS_{S\times W(T)}$. If either of the effects produces a nonsignificant $F$, proceed to Step C. Otherwise accept $H_0$.

(C) Test $F_1$ and $F_2$. If both $F$ ratios produce a significant result, reject $H_0$; otherwise accept $H_0$.

The simulation experiments of Forster and Dickinson (1976) validated this multistage procedure. Moreover,

the procedure is clear and precise. Why has it not become standard procedure? We assume that the problems in testing the effects of S×T and W(T) when missing data occur have severely restricted the use of their procedure. There are no easy ways of "repairing" the resulting empty cells. Perhaps this was a good reason to forget about Step B and to remember Steps A and C only. As was pointed out earlier, Step A has disappeared as well in the course of time, leaving the scene to Step C.

Despite the good results of the F&D procedure in their Monte Carlo studies, we suggest that an even better multistage procedure can be set up. The following considerations should play a role. $F_1$ and $F_2$ should be applied separately, depending on the properties of the data. Here the variance components $\sigma^2_{S\times T}$ and $\sigma^2_{W(T)}$ appear again. Table 2 substantiates the appropriateness of $F_2$ when the variance component $\sigma^2_{S\times T}$ is absent and the appropriateness of $F_1$ when the variance component $\sigma^2_{W(T)}$ is absent. We attach more importance than Forster and Dickinson did to the test of the interaction effect of subject-by-type (S×T), both for statistical and methodological reasons. Thus, we suggest the following multistage procedure.

### Multistage Procedure of Rietveld and van Hout (Hereafter, R&H)

(A) Test the effects of S×T and W(T).

(B) If the effect S×T is not significant, test $F_2$.

(C) If the effect S×T is significant and the effect W(T) is not significant, test $F_1$.

(D) If both effects S×T and W(T) are significant, test $F'$.

This decision procedure is not meant to be applied automatically. If the effect S×T is significant, an acceptable explanation is required. If the effect W(T) is not significant, indicating that there is no item variability effect, an explanation is needed as well, since item variability is a typifying source of variation in psycholinguistic experiments. For obvious reasons, the suggestion to "look for an acceptable explanation" was not formally used in the simulations reported in this article. The reason we prefer $F'$ over min $F'$ is its less conservative behavior. In Table 3, an overview of the two multistage decision procedures is given that illuminates the differences. Both of the procedures especially agree on the crucial role played by the two variance components $\sigma^2_{S\times T}$ and $\sigma^2_{W(T)}$.

## STATISTICAL SIMULATIONS

### Parameters

In the simulation experiments reported below, we wanted to assess the behavior of the four $F$ statistics that are currently in use when the results of repeated measures designs are reported—namely $F_1$, $F_2$, $F'$, min $F'$—and three decision procedures: the conventional $F_1$ & $F_2$, the F&D procedure, and the R&H procedure that we propose. The basic design comprised two within-subjects factors (type, with three levels, and word nested within type, 10 words [= items] per level of type) and 30 subjects. Thus, a complete data set comprised 900 data points, which functioned as simulated reaction time measurements. We assessed the behavior of these statistics as a function of three parameters:

(1) *Effect size* of the factor type, with values 0, 0, 0 (no effect), and −15, 0, +15 (small effect);

(2) Presence or absence of the *interaction effect* subject-by-type: $\sigma^2_{S\times T}$;

(3) Degree of *dependency* ("common variance") between words in different types (in this way, the matching of words between different types was simulated).

Three additional aspects will be discussed. The first one is the presence or absence of skewness in the distribution from which the words were drawn; the second, the impact of the effect size of the item variability; and the third, the way missing data can be handled.

### Data Generation

In order to estimate probabilities of Type I and Type II errors associated with different statistics and different characteristics of data sets, these data sets were generated by a FORTRAN program with routines from the Library Mark 18, made available by the Numerical Algorithms Group Ltd., Oxford, U.K.,[3] implemented on an IBM RS6000-R50 computer. For each condition, 10,000 data sets were generated. Since there were 2 (effect size) × 2 (presence/absence of the interaction S×T) × 11 (degree of dependency between words in different types: varied in steps of 10% between 0% and 100%) conditions, the results reported here are based on 44 × 10,000 = 44,000 data sets.

The model equation for the repeated measures design we used is

$$X_{ijk} = \mu + \alpha_i + \pi_j + \alpha\pi_{ij} + \beta_{k(i)} + \varepsilon_{ijk}, \qquad (1)$$

in which

$\mu =$ overall population mean (630 msec);

$\pi_j =$ the effect of subject $\pi_j$, random factor sampled from $N(0, 70)$, with a variance of $\sigma^2_S$;

$\alpha_i =$ the fixed effect of the three types $i$, with two options:
Effect (0, 0, 0) = 0,
Effect (−15, 0, +15) = 15

$\alpha\pi_{ij} =$ the interaction between subjects and type, with values set at 0 (S×T = 0) or taken from $N(0, 35)$ (S×T = 35), and variance $\sigma^2_{S\times T}$;

**Table 3**
**Overview of the Multistage Decision Procedures of F&D (Forster & Dickinson, 1976) and R&H (Rietveld & van Hout) in Relation to the Presence (= 1) or Absence (= 0) of the Variance Components $\sigma^2_{W(T)}$ and $\sigma^2_{S\times T}$**

| $\sigma^2_{W(T)}$ | $\sigma^2_{S\times T}$ | F&D | R&H |
|---|---|---|---|
| 0 | 0 | min $F'$ | $F_2$ |
| 0 | 1 | min $F'$; if not significant, $F_1$ & $F_2$ | $F_1$ |
| 1 | 0 | min $F'$; if not significant, $F_1$ & $F_2$ | $F_2$ |
| 1 | 1 | min $F'$ | $F'$ |

$\beta_{k(i)}$ = the effect of the $k$ words nested within $i$ types, random factor sampled from $N(0, 35)$—this is the "item variability," with variance $\sigma^2_{W(T)}$;

$\varepsilon_{ijk}$ = the error component, encompassing random error $\sigma^2_\varepsilon$ and the interaction $\sigma^2_{S\times W(T)}$ sampled from $N(0, 105)$.

Both the effects of the subjects and of the nested words were estimated on the basis of the variance components calculated for a fairly common data set with reaction times (Schreuder, Burani, & Baaijen, 2003).

**Common Variance**

How can the research practice of set matching be simulated? Wickens and Keppel (1983, p. 305) characterized item matching as matching samples on a number of relevant dimensions. They simulated it by stratifying word materials before selection. The distribution of words was divided into a number of strata or regions (= $b$). When $b = 2$, there was a region with words below the population median and a region above it. From each region a sample of $w$ words was randomly chosen. In essence, they added a blocking factor to the experimental design, varying the number of blocks (= $b$) from which words were sampled (the number of words = $w$) holding $b \times w$ (= 12) constant. The introduction of more blocks means a higher degree of matching. Next, the statistical analysis was performed without the blocking factor. They were interested "in the consequences of the relatively common practice of balancing samples of materials in the design of experiments without formally treating any blocking factor in the statistical analyis" (Wickens & Keppel, 1983, p. 306). The appropriate analysis was used ($F'$ for the design in which the blocking factor was included) as a reference point for the conventional $F_1$ and $F'$ analyses (the latter two were computed without the blocking factor). Their outcomes show that an increase in the number of blocks enhances the statistical performance of $F_1$ in a highly effective way (by reducing the Type I error rate). When blocking is ignored, $F'$ substantially lacks power.

We implemented a comparable procedure for set matching, but we opted for the explicit control over the common variance between sets of words. Our implementation is based on partially matching the sets involved at word level. *Partial* matching means that a specific degree of similarity is assumed between specific words in the sets. This degree of similarity is operationalized by assuming a common variance component between words and a unique component for each of the items. An example can easily demonstrate what partial matching is. Suppose we have a set of nouns and verbs and both sets have the same number of words. Each item in the noun set partially matches one item in the verb set. In fact we apply word-by-word matching. The word matching is partial, however, and in the final result at the set level no information is left with regard to which particular nouns were matched with which verbs. Word $n$ in the noun set is matched with word $n$ in the verb set as follows:

$$\text{word } n = w_1 \times (\text{sample value from common distribution}) + w_2 \times (\text{sample value from word-specific distribution}). \quad (2)$$

In our simulations, both random sample values are obtained from populations with the same mean (= 0) and the same standard deviation. The two samples are weighed by $w_1$ and $w_2$. The word $n$ gets the same mean and standard deviation, when the following relationship holds between the two weights (see Rietveld, van Hout, & Ernestus, 2004):

$$w_1^2 + w_2^2 = 1. \quad (3)$$

The degree of matching increases by taking higher $w_1$ values. For example, if we take a value of .80 for $w_1$, $w_2$ has the value .60. This results in a common variance of 64% (.80 × .80) and a word-specific variance of 36% (.60 × .60) of the total variance.

The procedure of partial matching has two obvious advantages over the blocking method applied by Wickens and Keppel (1983). They split up their population of words in a varying number of layers, but the number of layers did not have a transparent statistical interpretation. Our matching procedure has a statistical interpretation in terms of common variance. Second, the blocking method as applied by Wickens and Keppel systematically uses the complete range of variation in the population (they excluded only the extreme tails). This automatically means that (most of) the blocks can be reconstructed from the sample values of the selected words. In partial matching, the way the population variation is covered varies, which matches better the practice of word selection in psycholinguistic research, in which selection is often more determined by available words than by coverage of the full range between easier and more difficult words in terms of reaction time.

**Skewness**

The distributions from which the words were sampled were skewed in order to create ecologically valid conditions; skewness—$[\Sigma(x - \mu)^3/n]/\sigma^3$—was set at a value of 1, a value estimated on the basis of a data set used in Schreuder, Burani, and Baaijen (2003).[4]

**Missing Data**

Most of our simulations were carried out on complete data sets (30 subjects, 3 word types with 10 words nested within types, resulting in 900 "observations" in total), but additional simulations were carried out on similar data sets with 90 missing data each (10%). The missing data were obtained as follows: In each generation of a data set, two uniform distributions (ranging from 1 to 30) were sampled, each time generating the tuple $(i, j)$; this tuple determined the cell in the $z$-transformed data matrix, with $z$ values being calculated per word. In the next step, a value $P_u$ drawn from a uniform distribution ranging from .00 to 1.00 was compared with the left tail probability of $z(i, j)$. If $P_u < P[\text{left of } z(i, j)]$, the corresponding cell was given the code of "missing value." This procedure favored higher values being labeled as missing values, a process quite similar to what is found in real experiments with reaction times (the missings can be labeled *nonignorable* according to Allison, 2002). There are many imputation procedures, all with their own pros and cons

(see Allison, 2002). We used the "hot deck" procedure, also mentioned by Forster and Dickinson (1976) in their study of repeated measures designs in psycholinguistics. Of course, there might be some debate on this choice. We decided to use this procedure because it does justice to the randomness of both subjects and items (words), and because it was relatively easy to implement this procedure in our simulations. Our version of this procedure consists of randomly selecting—per subject and per type (treatment), with replacement—observed values to fill in the missing data. Conventionally, $F_1$ is calculated on the mean values per subject, pooled over words, and $F_2$ is calculated on the mean values per word, obtained by pooling over subjects.[5] The procedure was assessed by inspection of the Type I errors obtained when all requirements of the ANOVA procedure were met.

### Statistics

The outcomes of the simulations are evaluated on the basis of four statistics and three decision procedures.

$F_1$. Calculated either on the basis of a complete design in which words constituted a fixed factor and subjects a random factor (imputed data sets) or on the pooled data sets. As long as there are no missing data, $F_1$ calculated on the basis of means obtained by pooling over words is equal to the $F_1$ for the complete data set.

$F_2$. Calculated on the data set obtained by pooling over subjects, which was subsequently analyzed with a one-way analysis of variance. As long as there are no missing data, $F_2$ calculated on the basis of means obtained by pooling over subjects is equal to $F_2$ for the complete data set.

$F'$. Calculated for complete sets (see the formula below); obviously, this statistic cannot be calculated if only mean values of items are available; $F'$ is either based on complete data sets or sets in which missing values are replaced by imputation.

The $F'$ ratio is calculated as follows:

$$F' = \frac{MS_{\mathrm{T}} + MS_{\mathrm{S\times W(T)}}}{MS_{\mathrm{S\times T}} + MS_{\mathrm{W(T)}}}, \tag{4}$$

in which T = type of words, W(T) = word within type of words, and S = subject, with

$$df_1 = (MS_{\mathrm{T}} + MS_{\mathrm{S\times W(T)}})^2/(MS_{\mathrm{T}}^2/df_{\mathrm{T}} + MS_{\mathrm{S\times W(T)}}^2/df_{\mathrm{S\times W(T)}})$$

$$df_2 = (MS_{\mathrm{S\times T}} + MS_{\mathrm{W(T)}})^2/(MS_{\mathrm{S\times T}}^2/df_{\mathrm{S\times T}} + MS_{\mathrm{W(T)}}^2/df_{\mathrm{W(T)}}).$$

This corresponds to Equation 5, located at the bottom of the page, in which $p$ = number of types, $q$ = number of words at each level $\mathrm{T}_j$, and $r$ = number of subjects.

*min F′*. Minimum $F'$, calculated as $(F_1 \times F_2)/(F_1 + F_2)$. The degrees of freedom of the numerator are calculated as $df_2$ for $F'$; the same result can be obtained by calculating $(F_1 \times F_2)^2/(F_1^2/df_{\mathrm{W(T)}} + F_2^2/df_{\mathrm{S\times T}})$ (see Clark, 1973, pp. 356–357).

$F_1$ & $F_2$. The conventional decision procedure: Both $F_1$ and $F_2$ have to be significant to declare the factor type significant.

*F&D*. The procedure suggested by Forster and Dickinson (1976).

*R&H*. The procedure suggested by the authors (Rietveld & van Hout).

Moreover, we tested systematically the significance of the two effects that we manipulated in our simulations, namely S×T and W(T).

## RESULTS

### The Three Main Parameters

In the following, we present figures that contain the relative frequencies of occurrence of significant values (the significance level was set at .05) of the $F$ statistics and decision procedures. As explained earlier skewed data were used. Skewness of the distribution from which words are sampled, does not affect the $F$ ratios and associated $p$ levels substantially (see Kirk, 1995, p. 99). However, since the assumption that words are sampled from skewed distributions is more in agreement with reality than the assumption that they are taken from normal distributions, we have used skewed distributions throughout the simulations described below.

The simulations are based on three parameters. The first two are the absence/presence of an effect in the population and the absence/presence of the interaction S×T. Crossing these two parameters gives the four basic configurations in which the simulation results are presented in Figure 1. The third parameter is the degree of common variance which is varied between 0% and 100% in steps of 10%.

Our first simulations were carried out under the condition of no effect and no S×T interaction. Figure 1A gives the Type I errors of the different $F$ statistics and decision procedures.

The most conspicuous statistic in Figure 1A is $F_1$. As expected, this statistic turns out to be far too liberal, especially when the degree of common variance is low. When words is really a random factor (common variance = 0), the Type I error is even almost 50%. The Type I error rate of the $F_1$ drops quickly with increasing common variance, but it is only from a level of common variance of 80% or onward that the error rate reaches a level of about 5%. The remaining six statistics are not too liberal. They all perform at the 5% level or less. $F_2$ and R&H remain fairly constant at a 5% level. F&D starts at the 5% level, but its error rate clearly drops with increasing common variance. From a common variance of 50% onward, it gets the same low error rates as min $F'$. $F_1$ & $F_2$ and $F'$ show a drop in error rate as well with increasing common variance, but their drop is less steep.

$$F' = \frac{\sigma_\varepsilon^2 + \sigma_{\mathrm{S\times W(T)}}^2 + q\sigma_{\mathrm{S\times T}}^2 + \sigma_{\mathrm{W(T)}}^2 + qr\sigma_{\mathrm{T}}^2 + \sigma_\varepsilon^2 + \sigma_{\mathrm{S\times \mathbf{W}(T)}}^2}{\sigma_\varepsilon^2 + \sigma_{\mathrm{S\times W(T)}}^2 + q\sigma_{\mathrm{S\times T}}^2 + \sigma_\varepsilon^2 + \sigma_{\mathrm{S\times W(T)}}^2 + r\sigma_{\mathrm{W(\mathbf{T})}}^2} \tag{5}$$

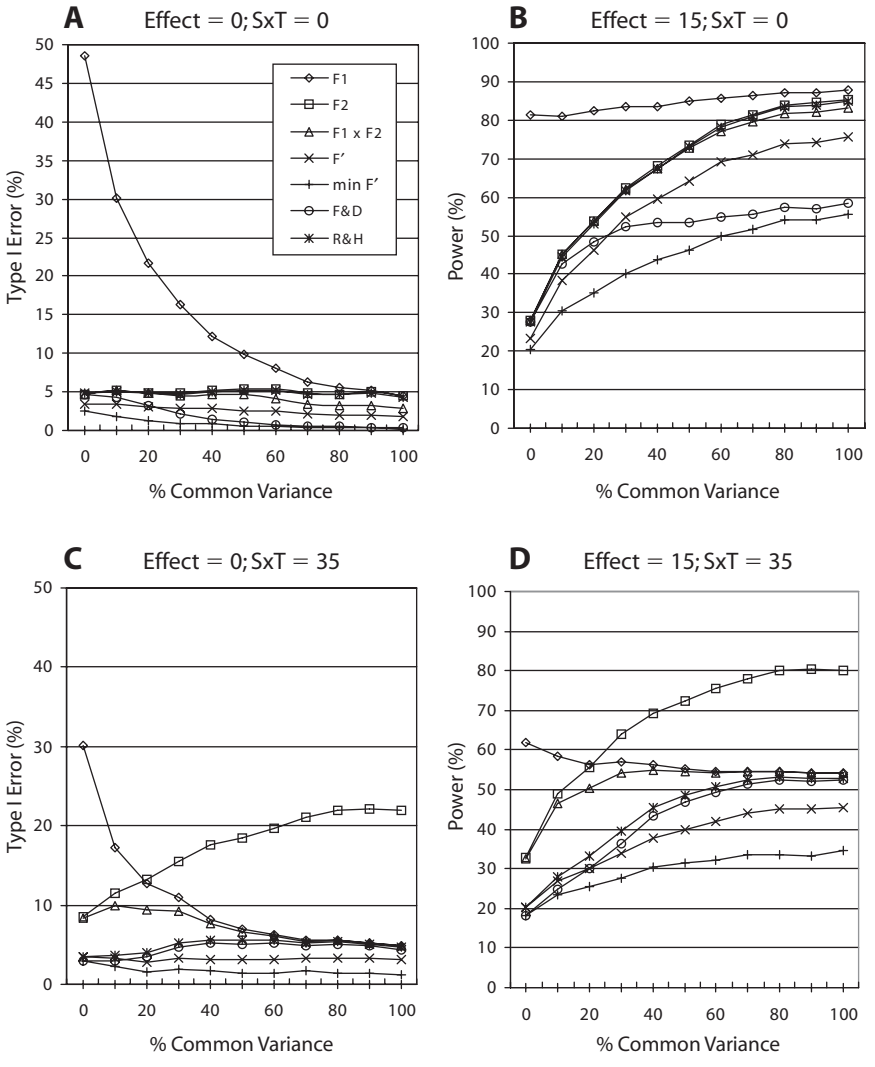


**Figure 1. Type I error rates (%) and power rates (%) for four *F* statistics and three decision procedures as a function of $\sigma_{S\times T}$ (either S×T = 0 or S×T = 35) and the degree of common variance; power rates are based on effect (−15, 0, 15).**

Figure 1B gives the power ratings for the statistics involved, when an effect is present, in combination with the absence of an S×T effect. All lines in this figure rise, which means that higher levels of common variance yield higher power ratings for all statistical measures involved, including $F'$ and min $F'$. $F_1$ has the highest power ratings, but we already saw in Figure 1A that the Type I error rates are much too high. Min $F'$ and F&D lag behind in power, min $F'$ over the whole common variance range, F&D when the common variance gets more substantial. The two best performing statistics are $F_2$ and R&H. This outcome confirms that $F_2$ is an excellent statistic, insensitive to the fixed-random distinction, as long as there is no S×T interaction. In that situation, the R&H procedure relies on the qualities of $F_2$.

The performances of $F_1$ and $F_2$ in Figures 1C and 1D are remarkable. The two figures cover the situation in which an interaction effect S×T is present. They show that the statistics involved behave differently than in the situation in which the S×T interaction is absent. The error rates in Figure 1C make clear, once again, that $F_1$ is too liberal, but its behavior improves as a function of increasing common variance. As expected, $F_1$ is perfect when the word effect reaches the level of fixedness. $F_2$ has the inverse problem: Its error rate rapidly inflates at higher levels of common variance. The consequence is that the $F_1$ & $F_2$ combination is too liberal as well, because both *F*s are too liberal in overlapping areas. $F_1$ & $F_2$ only gets an acceptable error rate at higher levels of common variance (70% and higher). Min $F'$ and $F'$ turn out to be conservative over the whole range of common variance. Both F&D and R&H perform well with a slight tendency to be conservative at low levels of common variance.

Figure 1D shows again the remarkable $F_1$ versus $F_2$ pattern that we found in Figure 1C. Because of their high Type I error rates, their power levels are of no interest here. F&D and R&H show the best power results, especially R&H when the whole range of common variances

**Table 4**
**Type I Errors of *F* Statistics and Decision Procedures As a Function of Common Variance and the Absence/Presence (0/1) of the Interaction S×T, Effect Size = 0**

| Common Variance | $\sigma^2_{S\times T}$ | $F_1$ | $F_2$ | $F_1$ & $F_2$ | min $F'$ | $F'$ | F&D | R&H |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | + | OK | OK | − | OK(−) | OK | OK |
| 20 | 0 | + | OK | OK | − | OK(−) | OK(−) | OK |
| 40 | 0 | + | OK | OK | − | − | − | OK |
| 60 | 0 | + | OK | OK | − | − | − | OK |
| 80 | 0 | OK | OK | OK(−) | − | − | − | OK |
| 100 | 0 | OK | OK | − | − | − | − | OK |
| 0 | 1 | + | + | + | − | OK(−) | − | OK(−) |
| 20 | 1 | + | + | + | − | OK(−) | OK(−) | OK |
| 40 | 1 | + | + | + | − | OK(−) | OK | OK |
| 60 | 1 | OK(+) | + | OK(+) | − | OK(−) | OK | OK |
| 80 | 1 | OK | + | OK | − | OK(−) | OK | OK |
| 100 | 1 | OK | + | OK | − | OK(−) | OK | OK |

Note—OK, error rate of about 5%; −, error rate is too low (<3%); OK(−), tendency to be too low (between 3% and 4%); +, error is too high (>7%); OK(+), tendency to be too high (between 6% and 7%); F&D, the decision procedure suggested by Forster and Dickinson (1976); R&H, the procedure suggested by Rietveld and van Hout.

is taken into account. The power rates in Figure 1D are lower than in Figure 1B, which is the consequence of the presence of the interaction effect S×T.

In Tables 4 and 5, we summarize our statistical findings by classifying the performances of the statistics involved. In Table 4 labels are assigned to the behavior of $F_1$, $F_2$, $F_1$ & $F_2$, min $F'$, $F'$, F&D, and R&H with respect to the Type I error. The relevant dimensions are again the amount of common variance and the absence or presence of the S×T interaction. The labels for the *F* statistics and the decision procedures are based on the outcomes presented in Figure 1. A plus sign in Table 4 means a Type I error value that is too high; this is a direct disqualification and, in fact, dismisses the associated statistic or decison rule as a serious candidate. $F_1$, $F_2$, and $F_1$ & $F_2$ fail to meet the Type I error requirements, and they particularly fail when subjects and types interact. Their performance substantially improves when common variance is high. A minus sign in Table 4 should not be seen as a negative qualification. Although a low Type I error value must have immediate implications for the power of the associated statistic or decision procedure, this does not matter in this table.

The power of the test statistics and decision procedures involved can be evaluated on the basis of the outcomes in Figures 1B and 1D. In Table 5, the results are summarized for the effect size of 15. Again, the relevant dimensions are the amount of common variance and the absence or presence of the interaction S×T. A plus for a specific cell in Table 4 was a knock-out condition for the corresponding cell in Table 5; this is marked by the label "No." The remaining cells contain their rank order in power row-wise. A higher rank means more power.

**Table 5**
**Power of *F* Statistics and Decision Procedures As a Function of Common Variance and the Absence/Presence (0/1) of the Interaction S×T, Effect Size = 15**

| Common Variance | $\sigma^2_{S\times T}$ | $F_1$ | $F_2$ | $F_1$ & $F_2$ | min $F'$ | $F'$ | F&D | R&H |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | No | 2.5 | 2.5 | 6 | 5 | 2.5 | 2.5 |
| 20 | 0 | No | 2 | 2 | 6 | 5 | 4 | 2 |
| 40 | 0 | No | 2 | 2 | 6 | 4 | 5 | 2 |
| 60 | 0 | No | 2 | 2 | 6 | 4 | 5 | 2 |
| 80 | 0 | 1 | 3 | 3 | 7 | 5 | 6 | 3 |
| 100 | 0 | 1 | 3 | 3 | 7 | 5 | 6 | 3 |
| Total | 0 | − | 14.5 | 14.5 | 38 | 28 | 18.5 | 14.5 |
| 0 | 1 | No | No | No | 3.5 | 1.5 | 3.5 | 1.5 |
| 20 | 1 | No | No | No | 4 | 2.5 | 2.5 | 1 |
| 40 | 1 | No | No | No | 4 | 3 | 1.5 | 1.5 |
| 60 | 1 | 1.5 | No | 1.5 | 6 | 5 | 3.5 | 3.5 |
| 80 | 1 | 2.5 | No | 2.5 | 6 | 5 | 2.5 | 2.5 |
| 100 | 1 | 2.5 | No | 2.5 | 6 | 5 | 2.5 | 2.5 |
| Total | 1 | − | − | − | 29.5 | 22 | 16 | 12.5 |

Note—Cells marked with a "+" in Table 4 referred to decisions with an error too high, and they are not considered here for power ranking; in this table, they are marked "No." The power of the *F* statistics is ranked row-wise, with a higher rank (= a lower number) indicating a higher power; differences of 2% or less are considered to be ties. F&D, the decision procedure of Forster and Dickinson (1976); R&H, the decision procedure of Rietveld and van Hout.

Table 5 shows that $F_2$ and $F_1$ & $F_2$ have high rankings when they are not excluded from the ranking procedure because of unacceptable Type I errors. The four statistical measures that are not marked by a "No" label (min $F'$, $F'$, F&D, and R&H) can be evaluated by comparing their total "power ratings" (obtained by summing their ranks). In the case of no interaction, R&H clearly has the best results, with a total power rating of 14.5. The totals of min $F'$, $F'$, and F&D indicate their conservative nature. In the case of the presence of an S×T interaction effect, both F&D and R&H have high rankings, with those of R&H being slightly better.

### Increasing Word Variability

An important question is how robust the results are. An essential aspect is the impact of word variability, which was set at 35 in the simulations we have presented thus far ($\sigma_{W(T)} = 35$). Word variability is important because the degree of common variance and the degree of word variability are interrelated phenomena. More common variance will reduce the word variability found in data sets. One could imagine that the results found are determined strongly by the relatively low level of word variability we used. Therefore, we carried out a series of simulations with a common variance of 60%, with word variability set at 105 ($\sigma_{W(T)} = 105$). The results are given in Figures 2A to 2D, which reflect again the four basic conditions we distinguished.

Figures 2A and 2C give the Type I error rates of the statistics and decision procedures for word variabilities of 35 and 105. The statistics and decision procedures do not exhibit much sensitivity to the two conditions of word variability, except for $F_1$ and $F_2$. We found a similar sensitivity to differing degrees of common variance. The other statistics and the two decision procedures seem fairly robust.

Figures 2B and 2D make clear that higher word variability leads to lower power. The relative positions taken by the statistical measures do not really change. If we rule out $F_1$, $F_2$, and $F_1$ & $F_2$ because of their high Type I error rates, R&H is still the best choice, as well as in cases of higher word variability.

### Missing Data

Missing data is an important topic in our context, because it seems to have been a factor contributing to the

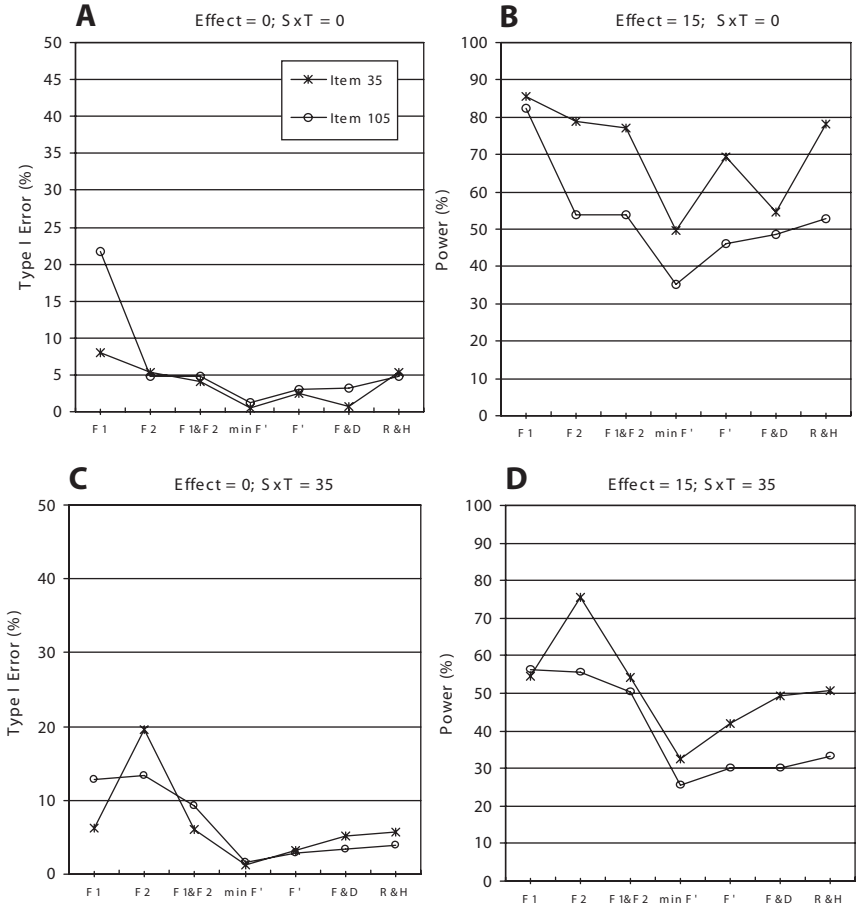


**Figure 2. Type I error rates (%) and power rates (%) for four *F* statistics and three decision procedures as a function of item variability ($\sigma_{W(T)} = 35$ vs. $\sigma_{W(T)} = 105$) and $\sigma_{S\times T}$ (either S×T = 0 or S×T = 35); power rates are based on effect (−15, 0, 15); common variance is 60%.**

existing preference for $F_1$ and $F_2$. However, we do not know how robust our statistical indices are when missing data occur. As said earlier, we also generated data sets with 10% missing values. We opted for a procedure that favored higher values of simulated reaction times to be missing. For $F_1$, $F_2$, $F_1$ & $F_2$, and min $F'$, the pooled procedures were used, disregarding the missing values, as is common practice in psycholinguistic research. For $F'$, F&D, and R&H, missing data were imputed on the basis of the hot deck procedure sketched earlier. The degrees of freedom were adjusted accordingly.

No important differences were found for the four $F$ statistics and the three decison procedures with respect to Type I errors, neither in the S×T = 0 condition nor in the S×T = 35 condition. Both pooling and imputation return Type I error rates that are somewhat lower compared to the situation where the data sets are complete, implying that the outcomes in case of missing values are somewhat more conservative. At the same time, this conclusion implies that the imputation procedure produces acceptable results.

For the two conditions in which an effect is present, we would expect a loss of power, since missing data entail loss of information in the data set. This is exactly what we observed in our simulations. There is another trend: The lower the common variance, the larger the losses in power. The losses in power relative to the nonmissing conditions vary between 0% and 10%, but these losses never change the general patterns we found for the simulations with nonmissing data. We may conclude that the simple hot deck procedure performs well and can be used when researchers opt for $F'$ or the decision procedures F&D and R&H. From a methodological point of view, the hot deck procedure is to be preferred over the pooling procedures.

## CONCLUSIONS

The simulation studies on the behavior of a range of statistics and decision procedures used in the analysis of repeated measures designs we have reported here, confirm a number of results and advices which have been mentioned in the literature since Clark's (1973) article on this subject. Our simulations were carried out in the framework of a simple, frequently used design with repeated measures, with two within-subjects factors ("type of words" and "word nested within type"). The relevance of the following points was confirmed in our simulations:

- When the nested factor word is a random factor, $F_1$ is not a suitable statistic because the associated Type I error is much too high. This effect is strengthened by higher word variability.
- The presence of the interaction of subject-by-type increases Type I error of $F_2$, which means that it is no longer a suitable statistic in this condition.
- Type I error associated with $F'$ and min $F'$ is low, at the expense of quite a low power.

We tried to put these points into a more general framework. First of all, we dealt with the interaction S×T and the W(T) effect in a systematic way in order to grasp their roles in the statistical evaluation of the type effect. We introduced a method of simulating set matching in language data (words), producing a factor that does not have an obvious fixed or random status. This method is based on the concept of common variance that makes it possible to define a scale between random and fixed.

We would like to point out here that the interaction S×T is not only a complicating factor from a technical and statistical point of view (its presence makes $F_2$ a useless test statistic, as was already pointed out by Forster & Dickinson, 1976). Textbooks on statistics remind us that the presence of an interaction effect sheds doubts on the relevance of the associated main effects. This is especially the case when the interaction is of the disordinal type (see Rietveld & van Hout, 2005). Interaction in our type of design means that subjects exhibit differential reactions to levels of the factor type; in the case of a disordinal interaction, they even show opposite patterns. Although this is common knowledge, it is amazing that a discussion of this interaction is absent in the psycholinguistic literature. We suggest that in psycholinguistic research more attention should be paid to differences between subjects, especially (so we would emphasize) when these differences have specific patterns related to the primary effects in the design such as type (see Baaijen, Tweedie, & Schreuder, 2002).[6]

Assessing the significance of the variance components S×T and W(T) is only possible—with the currently available statistical software for analysis of variance—if there are no missing data. When there are missing data, which is nearly always the case in psycholinguistic experiments, they have to be imputed. We found that the hot deck procedure is suitable for this purpose, both with low and high degrees of common variance between words. The hot deck procedure yields results similar to the ones obtained with $F_1$ and $F_2$ when missing values are ignored. The last aspect we want to mention is the assumption that many designs with repeated measures cannot be binarily categorized as designs with the factor word as either fixed or random. On the basis of the concept of "set matching," we claim that using a scale of "randomness" instead of the "random/fixed" dichotomy offers a more realistic approach to this problem. This scale was implemented in our simulations on the basis of two components, "common" variance and "word-specific" variance. It showed quite clearly the effects of the position of the factor word on this scale on Type I error and power.

All these considerations seem to be in favor of our decision procedure (R&H), which resembles the one suggested by Forster and Dickinson (1976), but differs in some essential respects. Relevant differences are that neither $F_1$ & $F_2$ nor min $F'$ are components of our decision procedure (see Table 3). The advantages and disadvantages of the various statistics and decision procedures are summarized in Table 6, which may function as a kind of consumer's guide.[7]

Although our findings, summarized in Table 6, are clear, it is hard to predict whether researchers in the field of psycholinguistics will follow our suggestions. Nevertheless, it cannot be denied that four procedures warrant a cor-

**Table 6**
**Evaluation Table of the Seven *F* Statistics and Decison Rules, on the Basis of Type I Error and Power**

| Statistic/Decision Procedure | Type I Error | Power |
|---|---|---|
| $F_1$ | OK, *iff* common variance is high *and* word variance is restricted | OK |
| $F_2$ | OK, *iff* S×T is absent | OK |
| $F_1$ & $F_2$ | OK, *iff* S×T is absent, *or* *iff* common variance is high *and* word variance is restricted | OK |
| min $F'$ | OK | too conservative |
| $F'$ | OK | tendency to be too conservative |
| *F&D | OK | OK, slightly conservative |
| *R&H | OK | OK |

*Procedures adequate in all conditions.

rect Type I error level in all conditions: min $F'$, $F'$, and the decision procedures F&D and R&H. In light of our results, the conventional $F_1$ & $F_2$ procedure needs to be abandoned.

## AUTHOR NOTE

We are indebted to Rob Schreuder (Centre for Language Studies, Radboud University of Nijmegen) for his generous and thoughtful counsel on many points in this article. Correspondence related to this article may be sent to T. Rietveld, Department of Linguistics, Radboud University Nijmegen, Erasmusplein 1, 6525 HT Nijmegen, The Netherlands (e-mail: a.rietveld@let.ru.nl).

## REFERENCES

ALLISON, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage.

BAAYEN, R. H., TWEEDIE, F. J., & SCHREUDER, R. (2002). The subjects as a simple random effect fallacy: Subject variability and morphological family effects in the mental lexicon. *Brain & Language*, **81**, 55-65.

CLARK, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning & Verbal Behavior*, **12**, 335-359.

COLEMAN, E. B. (1979). Generalization effects vs. random effects: Is $\sigma^2_{SL}$ a source of Type 1 or Type 2 error? *Journal of Verbal Learning & Verbal Behavior*, **18**, 243-256.

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 Min $F'$. *Journal of Verbal Learning & Verbal Behavior*, **15**, 135-142.

GORNBEIN, J. A., LAZARO, C. G., & LITTLE, R. A. (1992). Incomplete data in repeated measures analysis. *Statistical Methods in Medical Research*, **1**, 275-295.

GUEORGUIEVA, R., & KRYSTAL, J. K. (2004). More over ANOVA: Progress in analyzing repeated measures data and its reflection in papers published in the *Archives of General Psychiatry*. *Archives of General Psychiatry*, **61**, 310-317.

HUYNH, H. S., & FELDT, L. S. (1976). Estimation of the box correction for degrees of freedom. *Journal of Educational Statistics*, **1**, 69-82.

KIRK, R. E. (1995). *Experimental designs: Procedures for the behavioral sciences* (3rd ed.). Belmont: Brooks/Cole.

LEVENE, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 278-292). Stanford, CA: Stanford University Press.

LITTLE, R. J. A., & RUBIN, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

MAXWELL, S. E., & BRAY, J. H. (1986). Robustness of the quasi *F* statistic to violations of sphericity. *Psychological Bulletin*, **99**, 416-421.

QUENÉ, H., & VAN DEN BERGH, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, **43**, 103-121.

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 & Language*, **41**, 416-426.

RIETVELD, T., & VAN HOUT, R. (2005). *Statistics in language research: Analysis of variance.* Berlin: Mouton de Gruyter.

RIETVELD, T., VAN HOUT, R., & ERNESTUS, M. (2004). Pitfalls in corpus linguistics. *Computers & the Humanities*, **38**, 343-362.

SCHREUDER, R., BURANI, C., & BAAIJEN, R. H. (2003). Parsing and semantic opacity. In E. M. H. Assink & D. Sandra (Eds.), *Reading complex words* (pp. 159-189). Dordrecht: Kluwer.

WICKENS, T. D., & KEPPEL, G. (1983). On the choice of design and of test statistics in the analysis of experiments with sampled materials. *Journal of Verbal Learning & Verbal Behavior*, **22**, 296-309.

WILCOX, R. R. (1987). New designs in analysis of variance. *Annual Review of Psychology*, **38**, 29-60.

WINER, B. J., BROWN, D. R., & MICHELS, K. M. (1991). *Statistical principles in experimental design*. New York: McGraw-Hill.

## NOTES

1. There are other problems involved in the use of $F_1$ and $F_2$ that are never mentioned in psycholinguistic research, but for which there are solutions in standard statistical packages. When we calculate *F*, we deal with a repeated measures design. In repeated measures designs the assumption of "sphericity" must be met, a condition for the structure of the covariance matrix of differences between the levels of the within-subjects factor(s). Huynh and Feldt's index $\varepsilon$ assesses the extent to which this assumption is not met. The next step is to use the value of $\varepsilon$ to adjust the degrees of freedom involved. If $\varepsilon$ is smaller than 1, the *df*s are reduced, making the *F* test more conservative but bringing its significance level closer to the nominal value. As mentioned, $F_2$ is calculated on the basis of a one-way design. This one-way design assumes equal variances in the cells, an assumption currently tested with Levene's test. If this assumption does not hold, an inflation of the Type I error may occur (Wilcox, 1987). Both diagnostic statistics, $\varepsilon$ (Huynh & Feldt, 1976) for $F_1$ and Levene's test (Levene, 1960) for $F_2$, are hardly ever reported.

2. This also holds for the Mixed Models procedure in SPSS, in which maximum likelihood estimates are used.

3. The following NAG routines were used: G01EAF (returns a one- or two-tail probability from the standard normal distribution), G05CAF (returns a pseudorandom number taken from a uniform distribution between 0 and 1), and G05FDF (generates a vector of pseudorandom numbers taken from a normal distribution with mean *a* and standard deviation *b*). Details on the routines can be obtained via the NAG library manual, which is available on the Web: www.nag.co.uk/numeric/FL/

manual/html/. Seeds were provided by the internal computer clock. Output data sets were tested for accuracy by inputting them into SPSS. The data sets turned out to reproduce the intended characteristics.

4. We thank Ton de Haan of the Department of Medical Informatics, Epidemiology and Statistics of Radboud University Nijmegen for providing us with the algorithm to implement skewness in our simulations.

5. Recent approaches to missing data are the maximum likelihood (ML) and estimation maximum likelihood (EM) methods; for SPSS, the latter is only available in the Windows version, and not in SPSS-Unix, the software we had to use for our simulations. The Winer, ML, and EM approaches all result in covariance matrices of the completed ("imputed") data set per condition which are maximally similar to approximations of the covariance matrices of the original (incomplete) data sets. An unwanted by-effect of these procedures, however, is the emergence of an S×T interaction, which will be declared significant on the basis of the high power of the associated $F$ ratio, with $df_1 = 58$ and $df_2 = 693$ (for 90 missing data). In simulations with 10,000 data sets with no effect for type, with an S×T interaction, and with the $k$ words nested within the $i$ types sampled from $N(0, 35)$ and $\varepsilon_{ijk}$ (the error component) sampled from $N(0, 105)$, 35.76% significant $F$ ratios (at the 5% level) for this interaction (mean $F$ ratio is 1.27, with $SD$ of 0.25) were found for the Winer procedure, and for ML this percentage was about 10% lower: 25.48%, with mean $F$ ratio of 1.23 and quite a high $SD$ of 0.93. The low $F$ ratios suggest small effect sizes for S×T. Since the detection of an S×T interaction is important both for methodological reasons and for the assessment of $F_2$—which is quite sensitive, as our results will confirm, to the presence of this variance component—we decided to implement a version of an imputation procedure known as "hot deck" (see Little & Rubin, 1987; this procedure is also mentioned by Forster & Dickinson, 1976). Our version of this procedure consists of randomly selecting—per subject and per type (treatment), with replacement—observed values to fill in the missing data. Thus, the subject effect will be saved at the expense of the word effect; we think that this is in agreement with the purpose of a within-subjects design, in which the random factor word is seen to provide equivalent items, and subjects are assumed to behave with their own bias throughout the experiment. In simulations with the hot deck procedure, the percentage of incorrect significant S×T interactions dropped substantially, to 6.22%; the mean $F$ ratio was 1.02, with $SD = 0.20$. We thank Jan van Leeuwe (technical support group, Faculty of Social Sciences, Radboud University Nijmegen) for his comments on the approaches to missing data.

6. The role of the nested factor word, W(T), is somehow difficult to grasp conceptually. One of the primary reasons may be that it virtually conflates two variance (sub)components, $\sigma^2_{W\times T}$ and $\sigma^2_W$. Of course, these two components cannot be separated as long as words are nested within type, but we need these two components in order to understand what happens to this effect when set matching is applied. Increasing set matching implies a (serious) decrease in power in the detection of word variability. Another way of putting this is that set matching decreases the impact of word variability. Our simulations show that it is a successful strategy to use $F_1$ when W(T) is a nonsignificant effect.

7. Independent of the decision taken to use a particular statistic or decision procedure, the normal tests for assumptions underlying the statistics should be carried out: For $F_1$ it is Huynh–Feldt's test for sphericity, and for $F_2$, Levene's test of homogeneity. In an informal survey of articles published in authorative journals in the last 4 years, we have hardly ever seen reports of Huynh and Feldt's test for sphericity and Levene's test of homogeneity, which adds to our doubts about the validity of reported $F_1$s and/or $F_2$s.

(Manuscript received April 4, 2006;
revision accepted for publication December 11, 2006.)