

What exactly is random about random effects?

Matthias Siemer

Abstract

The central question of the paper is: When should stimuli be treated as random effects in random- or mixed-effects ANOVA? The conceptual starting point is the view of significance testing in experimental research as following the logic of randomization tests. Three designs are discussed: (1) Stimuli nested under treatment conditions, (2) stimuli and treatment conditions crossed, and (3) stimuli and treatment conditions counterbalanced. The treatment of stimuli as random effects in the first two designs is shown to be inadequate but is necessary in the third design.

Moreover, the analysis has implications for at least two additional topics: (1) The conceptualization of the significance test as a randomization test when no randomization takes place, e.g., in quasi-experiments and correlational studies, and (2) the validity of aggregating over subjects and stimuli (referred to as *aggregation validity*). Aggregation validity has to be distinguished from internal validity and is shown to be one aspect of the concept formerly known as external validity, albeit without appeal to some underlying population.

The present paper starts with the following methodological question: Under which circumstances is it reasonable to consider stimuli in experimental studies to be random samples drawn from an underlying population of stimuli. By implication these stimuli would have to be treated as random factors in a mixed- or random-effects ANOVA-model. The answer to this question has to consider the following basic question: What is the exact nature of the supposed *randomness* of random effects? The Discussion focuses on the methodological implications and presuppositions of this assumption of randomness rather than on technical problems of the random-effects model, like the distributions of *Quasi-F* statistics.

1 Arguments for the treatment of stimuli as random effects

It is established methodological standard in cognitive psychology to treat stimuli – e.g. words – as a random factor, just like subjects. Indeed, cognitive psychologists often carry out two distinct statistical analyses, one with subjects as random effects (and stimuli treated as fixed effects) and the other one with stimuli as random effects (and subjects treated as fixed effects). The rationale for this approach goes back to Coleman (1964) and Clark (1973). For instance, Coleman stated that

many studies of verbal behavior have little scientific point if their conclusions have to be restricted to the specific language materials that were used in the experiment. It has not been customary, however, to perform significance tests that permit *generalization* beyond the specific materials, and thus there is little evidence that such studies could be

successfully *replicated* if a different sample of language materials were used (1964, p. 216, *italics added*).

This methodological requirement has since been extended to other research areas (e.g. Fontenelle, Phillips & Lane, 1985; Hopkins, 1984; Kenny & Smith, 1980; Richter & Seay, 1987; Santa, Miller, & Shaw, 1979; Wickens & Keppel, 1983) and the statistical treatment of different experimenters (Bonge, Schuldt, & Harper, 1992; Scheirer & Geller, 1979). The arguments in favor of this methodological requirement are generally not substantially different from those of Coleman (1964). The following paragraph from Fontenelle, Phillips, and Lane (1985) may be taken as representing this view:

In conclusion, *generalizing* research findings is a major aim of all experimentation. Although experimenters as a rule are very careful to make sure that their results generalize to the population of subjects, the problem of generalizing to the population of stimuli had been neglected (p. 106, *italics added*).

Before the validity of these arguments can be discussed further, it is necessary to briefly present the widely held conceptualization of inferential statistics in experimental psychology as following the logic of randomization tests (rather than parametric or population based statistics).

2 The significance test as randomization test

Recent interpretations of the significance test do not focus on it as a means to infer to an underlying population – at least not in experimental contexts. Rather, the significance test is seen as following the logic of randomization tests (cf. Bredenkamp, 1980; Edgington, 1995; Erdfelder & Bredenkamp, 1994; Gadenne, 1984; Hager & Westermann, 1983). In randomization tests, the critical random process is not to draw a random sample from an underlying population distribution, but rather to randomly assign subjects (or stimuli) to experimental conditions. This random assignment is required to secure stochastic independence of *treatment* and *potentially confounding variables* (PCVs). PCVs designate all other factors that have an effect on the dependent variable/s in the sense of being (parts of) sufficient conditions for causal influence on the dependent variable/s. The latter reflects the fact that psychological hypotheses and explanations are essentially incomplete: They entail only one or few causes as parts of a complete causal network. Moreover, they are neither necessary nor sufficient in themselves but have to be supplied by other (background-) conditions to be effective (see Siemer, 1993 for a more detailed discussion of this property of psychological explanations). PCVs become *actually confounding* variables (ACVs) inasmuch as the stochastic independence of PCV and treatment does not hold. The randomization procedure is therefore required to secure the internal or *ceteris-paribus* ("other things equal") validity of an experiment (Hager & Westermann, 1983). If the treatment procedure itself induces confounding variables, randomization does not ensure internal validity, since the confounding takes place after randomization (e.g., Cook & Campbell, 1986).

According to this rationale, a randomization test informs about the probability of the empirical data (in terms of differences in central tendencies), under the assumption that they are exclusively the result of the random assignment procedure, that is, *chance*. Most importantly, the distribution of the test statistic is generated solely on the basis of the sample data.

To summarize, randomization is necessary to secure the internal validity of an experiment that tests causal hypotheses. Therefore, the randomization test answers

the question of how likely the results are, under the assumption that the treatment has no effect (H_0) and all effects are exclusively a consequence of the randomization procedure. Consequently, the randomization test (with appropriate α -level) protects from falsely accepting a causal hypothesis, and consequently increases the internal validity of an experiment. In particular, inferential statistics do not provide the (inductive) generalization from a sample to a population, that is, the "external validity" or expected probability to replicate the results of an experiment.

Therefore, the question of interest can be rephrased as: Under which circumstances does the treatment of stimuli as random effects raise the internal validity of a study? Note that in the present context ANOVA is treated as an approximation to randomization tests. This is possible because Monte-Carlo studies have shown that randomization tests and their parametric equivalents in most cases lead to very similar results (summarized in, e.g., Bredenkamp, 1980). The main reason for this approach is that ANOVA is a much more common method for the evaluation of experimental designs than randomization tests, at least for the designs considered. Therefore, from a *statistical* point of view, I will discuss random effects or mixed ANOVA models, whereas from a *conceptual* or methodological point of view, these ANOVAs will be treated as approximations to randomization tests.

3 Subjects as random factors

To illustrate this approach I will discuss a design that requires the treatment of *subjects* as random effects – both at a statistical and a conceptual level: A single factor repeated measurements design with two treatment levels. It is assumed, that sequence of the factor levels and subjects are randomly combined. The respective randomization test informs about the probability of empirical mean differences, given the H_0 that these differences are exclusively the result of the random combination of the sequence of the factor levels and the subjects. That is, the probability that the effect is the result of chance fluctuations in the reactions of the subjects across the two levels. The notion of "chance fluctuation" subsumes the variation of the influences of all PCVs at the two points of measurement. Consequently, the distribution of the test statistic is generated by the permutation of the combinations of reactions and treatment levels within subjects (e.g., Edgington, 1995).

In the ANOVA evaluation of this design, subjects are treated as random effects in a mixed-model approach. This is reflected by the fact that the F -ratio for the expected mean squares of treatment variability is compared to the expected mean squares of the interaction of treatment by subjects, consisting of the variance components σ_e^2 and $\sigma_{T \times S_u}^2$ (see Table 1). These variance components (which are confounded in case of only one observation per stimulus-subject combination) can be considered to be estimations of the influence of pure chance fluctuations across the points of measurement, since these fluctuations will cause differential treatment effects. Note however, that the variance component $\sigma_{T \times S_u}^2$ entails also "true" or manifest differential effectiveness of the treatment. As a consequence, the expected mean square will *overestimate* the amount of pure chance fluctuations across the points of measurement. Moreover, the hypothesis tested usually does not predict that the treatment has numerically *the same* effect on all subjects and therefore allows for *ordinal* treatment-by-subject interactions. In the present design, however, chance variation and ordinal treatment-by-subject interactions are not separable.

As a result, the mixed-model ANOVA resembles the randomization test account of chance fluctuations by comparing treatment variance to an upper bound estimation of pure chance fluctuations across the points of measurement (including error variance σ_e^2) in the critical F -value.

Based on these arguments, the revised question now is: Under which circum-

Table 1: Sources of Variance and Expected Mean Squares; Repeated Measurements Design

Source	df	E(MS)
T	$p - 1$	$\sigma_e^2 + \sigma_{T \times Su}^2 + n\sigma_T^2$
Su	$n - 1$	$\sigma_e^2 + p\sigma_{Su}^2$
$T \times Su$	$(n - 1)(p - 1)$	$\sigma_e^2 + \sigma_{T \times Su}^2$ (Residual)

Note. $T(p)$ = Treatment; $Su(n)$ = Subjects

stances does the treatment of stimuli as random effects serve the same purpose as the treatment of subjects in the present design does? Three different design are distinguished.

4 Stimuli nested under treatment conditions

In this common design, stimuli are nested under treatment conditions. As a result, both factors are confounded. Examples are studies in which different word classes (e.g. nouns vs. adjectives) serve as treatments. The linear model of an individual score x_{ijm} of subject m on stimulus j in treatment condition i is given by Equation (1).

$$x_{im} = \mu + \alpha_i + s_{j(i)} + p_m + \alpha p_{im} + s p_{mj(i)} + e_{m(ij)} \quad (1)$$

The terms of Equation (1) specify the potential sources of variability in the (quasi-) experiment. The term μ represents the overall mean and α_i , $s_{j(i)}$, and p_m are the treatment, stimulus and subject effects, respectively. (The grand mean and the treatment effect are designated by Greek letters to indicate that they are assumed fixed; the others are potentially random.) The treatment-by-subject interaction is expressed as αp_{im} . The quantity $s p_{mj(i)}$ is a particular stimulus-by-subject interaction, while $e_{m(ij)}$ is the random error associated with that particular subject-stimulus combination in that particular experiment. Table 2 shows the respective expected mean squares of the ANOVA models for the random- vs. fixed-effects model. In the random-effects model, the mean square of the treatment effect has an expected value that entails the stimulus variance-component (σ_T^2). Therefore, the appropriate error term to test the treatment effect is one that includes this source of error, resulting in a *Quasi-F*-ratio (e.g., Clark, 1973). In contrast, in the fixed-effects model, variability originating from stimulus variability is not considered.

In order to find the appropriate model one first has to analyze whether the hypotheses about the stimuli have the same nature as those previously discussed, that is: Do the critical hypotheses concern the populations of stimuli or are they essentially causal? Only if the stimulus hypotheses are actually hypotheses about populations it is necessary and reasonable seek for *statistical* generalization by means of random sampling of material.

Even looking at the research questions stated by the authors arguing in favor of the treatment of stimuli as random effects (e.g., Clark, 1973) makes clear, however, that the critical hypotheses are far from being hypotheses about central tendencies in clearly defined and closed populations of stimuli. Instead, the hypotheses concern the *causal properties* of certain features of the stimuli. The appeal to populations serves a completely different purpose. Consider, for instance, the following explanation of what Clark (1973) sees as the purpose of a "central tendency" hypothesis

Table 2: Sources of Variance and Expected Mean Squares; Repeated Measurements Design with Stimuli Nested Under Treatment-Conditions.

Source ^a	df	E(MS) ^b
T	$p - 1$	$\sigma_e^2 + \sigma_{\mathbf{St}(\mathbf{T}) \times \mathbf{Su}}^2 + q\sigma_{T \times Su}^2 + r\sigma_{\mathbf{St}(\mathbf{T})}^2 + qr\sigma_T^2$
St/T	$p(q - 1)$	$\sigma_e^2 + \sigma_{St(T) \times Su}^2 + r\sigma_{St(T)}^2$
Su	$r - 1$	$\sigma_e^2 + \sigma_{\mathbf{St}(\mathbf{T}) \times \mathbf{Su}}^2 + pq\sigma_{Su}^2$
$T \times Su$	$(p - 1)(r - 1)$	$\sigma_e^2 + \sigma_{\mathbf{St}(\mathbf{T}) \times \mathbf{Su}}^2 + q\sigma_{T \times Su}^2$
$St/T \times Su$	$p(p - 1)(r - 1)$	$\sigma_e^2 + \sigma_{St(T) \times Su}^2$ (Residual) ^c

a. T = Treatment (p), Su = Subjects (n), St = Stimuli (q)

b. Variance components typed in bold letters are parts of the random – but not fixed – effects model.

c. σ_e^2 and $\sigma_{St(T \times Su)}^2$ can not be estimated independently with only one observation per stimulus-subject combination.

(i.e., a population hypothesis about aggregates) in the context of a comparison of homographs vs. nonhomographs:

... homographs take longer to recognize than nonhomographs *all other things being equal*. Since it is impossible to find single homograph/nonhomograph pairs identical in all other possible factors – frequency, meaning, word length, spelling difficulty, and other undetermined factors – it is only possible to test the hypotheses by looking at the central tendencies (for example, the means) of homographs versus nonhomographs. (Clark, 1973, p. 352, *italics added*)

This paragraph makes perfectly clear, that the hypothesis of interest does not concern aggregates of stimuli in well defined populations. Rather, the very property of homography *itself* is assumed to causally influence identification times. This is exactly the reason why all other properties of the stimuli have to be controlled in order to realize the ceteris-paribus condition (and secure internal validity). More precisely, it is only necessary to control PCVs and not *all* other properties of the stimuli, because some of these properties have no causal effect. Why the ceteris-paribus condition is assumed to be true at the level of central tendencies in populations, rather than in samples, remains entirely unexplained, however. Indeed, this is true, only if a very restrictive additional assumption is met.

First of all, how is a conceptualization of the significance test as a randomization test in the present design possible at all? Obviously, the stimuli are not randomly assigned to conditions, to begin with. One could argue, however, that the test of significance informs about the probability of the empirically obtained differences in central tendencies, under the assumption that the stimuli come from *the same population*. If this assumption is met, random sampling has the same effects as random assignment to conditions. The assumption, that the stimuli are drawn from the same population (H_0) can in turn be reduced to two underlying assumptions:

1. The null hypothesis is true, that is, the stimuli come from the same population with respect to the treatment.
2. The stimuli come from the same population with respect to all PCVs.

The second assumption implies nothing less than the assumption of the validity of the ceteris-paribus condition at a population level. This assumption becomes necessary, because the ceteris paribus condition is not guaranteed by a randomization procedure (at least with respect to the distributions of the PCVs, e.g. Steyer,

1992). It has thus to be secured in a different way. However, this is a very strong *a-priori assumption*. As a result, the significance test informs only about the probability of the data, given the conjunction of the H_0 and the assumption that the ceteris-paribus condition is met at the population level. The internal validity of the inference is therefore secured, *if and only if* there exist no systematic differences (in terms of PCVs) between the kinds of stimuli in the population, that is, stimulus category and PCVs are stochastically independent. For a recent approach to falsify this assumption of *unconfoundedness* in the population with respect to single PCVs – referred to as *potential confounders* – see Steyer, Gabler, and Rucai (1995).

The same line of argument holds – *mutatis mutandis* – for other quasi-experimental designs or correlational studies in which randomization is not possible – at least as far as causal and not “true population hypotheses” are concerned (see the Conclusions). True population hypotheses, however, are commonly of limited theoretical value, since the major aim of (basic) science lies in the investigation of general causal mechanisms rather than in the description of incidental and local phenomena.

Usually, one tries to increase the ceteris-paribus validity in these cases by quasi-experimental control techniques like matching or the inclusion of PCVs either in the experimental design or in the regression or path analysis. However, matching with respect to stimuli seems to be more feasible than with respect to subjects, because the number of PCVs is more restricted. Indeed, if it were possible to reach a “perfect matching” (i.e., with respect to all PCVs), or to include all PCVs in a regression analysis (thus fulfilling the so called “closedness” condition), it would be possible to interpret a regression coefficient or a difference in central tendencies unambiguously in a causal manner.

For the present design this means that the stimulus variance that is not evoked by the treatment (σ_{St}^2) is reduced to zero. As this variance component is exactly the term that separates the random-effects from the fixed-effects model, this means that differences between the two models are reduced inasmuch as a matching of the stimuli is successful. Furthermore, as blocking factors (e.g., word length) are usually considered in the process of designing a study but not in the subsequent statistical analysis, the application of the random-effects model results in a smaller actual α -level than the nominal one (even if the statistical assumptions of the random-effects model, like true random sampling, are actually met; cf. Wickens & Keppel, 1983). Elimination of extreme material (i.e., truncation of the stimulus distributions) has the same effect (e.g., Cohen, 1976).

5 Stimuli and treatment conditions crossed

In this design stimuli and treatment are completely crossed, with subjects nested under treatment levels. In other words, the design is within stimuli and between subjects. The linear model for a single score x_{ijm} of subject m on stimulus j in treatment condition i is given by Equation (2):

$$x_{ijm} = \mu + \alpha_i + s_i + p_{m(i)} + \alpha s_{ij} + sp_{jm(i)} + e_{ijm} \quad (2)$$

In contrast to Equation (1), subjects instead of stimuli are nested under the treatment conditions. The resulting variance components of the models with stimuli treated as fixed versus random effects is presented in Table 3. The main difference between the random- and the fixed-effects model consists in the variance component $\sigma_{T \times St}^2$, that is, the interaction between stimuli and treatment. This interaction is included in the expected means of the treatment effect in the random- but not the fixed-effects model. Consequently, the appropriate *Quasi-F* ratio has to include the respective variance component in the error term (e.g., Hopkins, 1984). In contrast,

Table 3: Sources of Variance and Expected Mean Squares; Between-Subjects Design with Stimuli Crossed with Treatment-Conditions.

Source ^a	df	E(MS) ^b
<i>T</i>	$p - 1$	$\sigma_e^2 + \sigma_{\mathbf{St} \times \mathbf{Su} / \mathbf{T}}^2 + n\sigma_{\mathbf{T} \times \mathbf{St}}^2 + q\sigma_{\mathbf{Su} / \mathbf{T}}^2 + nq\sigma_T^2$
<i>St</i>	$q - 1$	$\sigma_e^2 + \sigma_{\mathbf{St} \times \mathbf{Su} / \mathbf{T}}^2 + np\sigma_{\mathbf{St}}^2$
<i>Su/T</i>	$p(n - 1)$	$\sigma_e^2 + \sigma_{\mathbf{St} \times \mathbf{Su} / \mathbf{T}}^2 + q\sigma_{\mathbf{Su} / \mathbf{T}}^2$
<i>T × St</i>	$(p - 1)(q - 1)$	$\sigma_e^2 + \sigma_{\mathbf{St} \times \mathbf{Su} / \mathbf{T}}^2 + n\sigma_{\mathbf{T} \times \mathbf{St}}^2$
<i>St × Su/T</i>	$p(q - 1)(n - 1)$	$\sigma_e^2 + \sigma_{\mathbf{St} \times \mathbf{Su} / \mathbf{T}}^2$ (Residual) ^c

a. T = Treatment (p), Su = Subjects (n), St = Stimuli (q)

b. Variance components typed in bold letters are parts of the random – but not fixed – effects model.

c. σ_e^2 and $\sigma_{\mathbf{St} \times \mathbf{Su} / \mathbf{T}}^2$ can not be estimated independently with only one observation per stimulus-subject combination.

the fixed-effects model resembles mathematically a oneway ANOVA for the aggregated scores of the stimuli, thus ignoring differential treatment influences on the stimuli.

The present design closely resembles the repeated measurements design with subjects as random effects. In this previously discussed design the treatment-by-subjects interaction variance component ($\sigma_{\mathbf{T} \times \mathbf{Su}}^2$) serves as an estimation of the amount of chance fluctuations in the reactions of the subjects – together with the true or manifest treatment-by-subjects interaction caused by "real" differential effectiveness of the treatment.

In contrast, with respect to stimuli this conclusion is not valid. Stimuli – as opposed to subjects which are essentially "open systems" – are physically identical at different points of measurement. As a consequence, it is not possible to have a reasonable notion of chance fluctuations with respect to stimuli in the same way as it is for subjects. Stated differently: The treatment-by-stimulus interaction ($\sigma_{\mathbf{T} \times \mathbf{St}}^2$) is completely manifest. Consequently, $\sigma_{\mathbf{T} \times \mathbf{St}}^2$ is an exclusive indicator of whether the treatment influences all stimuli in the same way and to the same extent. As a result, this measure has no implications for the question of internal validity. To answer the question of whether an observed effect can be causally related to the treatment variation at all or if it can also be explained by chance, it is irrelevant whether this effect is identical across all stimuli. The latter is a question of external or – in analogy to the concept of population- validity (Hager & Westermann, 1983) – *stimulus validity*. To accomplish a high level of stimulus validity one first of all has to secure that the stimuli investigated are actually a sample from the population of stimuli for which the tested hypothesis demands validity. This assumption is trivially met, if the theory tested does not restrict the population of stimuli to some subpopulation. The second aspect of stimulus validity is the question whether the treatment causally influences all stimuli in the same direction. That is, whether *disordinal* interactions exist between treatment and (subgroups of) stimuli. In analogy to the notion of "aptitude-treatment interaction", one could speak of *stimulus-treatment interactions*. Only with respect to this question $\sigma_{\mathbf{T} \times \mathbf{St}}^2$ becomes important. Therefore, $\sigma_{\mathbf{T} \times \mathbf{St}}^2$ should at least be reported descriptively. Additionally, one should investigate whether partitioning of the stimuli – according to some control variables – is possible and accounts for relevant proportions of $\sigma_{\mathbf{T} \times \mathbf{St}}^2$. Moreover, it might be helpful to report and compare different measures of effect size on the basis of intraclass correlations (also denoted as "generalizability" coefficients, e.g. Cronbach, Gleser, Nanda, & Rajaratnam, 1972). In effect, what is

considered here is whether the treatment has the same effect on all stimuli, that is, whether an aggregation across different stimuli leads to valid conclusions. If disordinal stimulus-treatment interactions exist, aggregation across subgroups of stimuli is no valid approach (see Iseler, 1996a, 1996b for a discussion of the deductive relation of single case and statistical aggregate hypotheses and the methodological and statistical implications). In order to allow for a more differentiated terminology concerning questions of validity in experimental research, this second aspect of stimulus validity – or population validity for that matter – might be referred to as *aggregation validity*. This kind of validity, however, is entirely different from the question of “pure” internal validity, that is, whether the treatment has causal effects *at all*.

6 Stimuli and treatment conditions counterbalanced

This design results, if stimuli and treatment levels can be combined freely, but each stimulus can be presented only once. In this design stimuli and treatment conditions are balanced by building two sets of stimuli and subjects (in case of two treatment levels) and crossing them. The treatment effect equals the interaction of the dummy-coded subject (A) and stimulus (B) sets (cf. Kenny & Smith, 1980). With respect to the two dummy variables subjects and stimuli are nested. Equation (3) shows the linear model for a single score x_{ijkm} .

$$x_{ijkm} = \mu + \alpha_i + \beta_j + s_{k(j)} + p_{m(i)} + \alpha\beta_{ij} + \alpha s_{ik(j)} + \beta p_{mj(i)} + \gamma p_{k(j)m(i)} + e_{m(ijk)} \quad (3)$$

In the counterbalanced design both, subjects as well as stimuli, are randomly assigned (in the present case to two groups), that is, stimuli and subjects are treated symmetrically. Consequently, the number of the subjects and the stimuli should be equal. If stimuli are treated as random effects the treatment effect entails the variance component $\sigma_{B \times St}^2$, just the same way as the treatment of subjects as random effects leads to the inclusion of the variance component $\sigma_{A \times St}^2$ (see Table 4). In both cases this variance component reflects an estimation of the amount of variance that can be attributed to the randomization of subjects and stimuli, respectively. This reflects the ANOVA approximation of a between subjects randomization test. As a consequence, from the perspective of the subjects as well as from that of the stimuli, an ANOVA in which both, subjects and stimuli, are treated as random effects is a valid conceptual approximation of a randomization test in which the distribution of the test statistic is calculated by permutating both stimuli and subjects. Therefore, inasmuch as tests of significance – in the interpretation of being randomization tests – raise the internal validity of an experiment, this is also true for the treatment of stimuli as random effects in the counterbalanced design.

One of the disadvantages of the random-effects model has not yet been discussed (because the application of the random-effects model did not prove to be advisable): The underlying variance components model is not as robust against violations of its statistical assumptions as the fixed-effects model, which is analyzable in the context of the General Linear Model. For instance, the random-effects model is less robust against violations of the normality assumption, since the central limit theorem is not applicable to variance components, as opposed to means (e.g., Scheffé, 1963). Moreover, the distributions of the *Quasi-F* statistics, the basis for tests of significance, cannot be derived analytically and have to be analyzed by means of Monte-Carlo studies (e.g., Maxwell & Bray, 1986; Santa, Miller, & Shaw, 1979). For example, in the present case, the resulting *Quasi-F* ratio is given by Equation (4), cf. Kenny and Smith (1980).

Table 4: Sources of Variance and Expected Mean Squares; Counterbalanced Design with Two Stimulus and Subjects Groups.

Source ^a	df	E(MS) ^b
<i>A</i>	1	$\sigma_e^2 + \sigma_{\mathbf{St} \times \mathbf{Su}}^2 + \mathbf{n}\sigma_{\mathbf{A} \times \mathbf{St}}^2 + qr\sigma_{\mathbf{Su}}^2 + qrn\sigma_T^2$
<i>B</i>	1	$\sigma_e^2 + \sigma_{\mathbf{St} \times \mathbf{Su}}^2 + r\sigma_{\mathbf{B} \times \mathbf{Su}}^2 + \mathbf{pn}\sigma_{\mathbf{St}}^2 + prn\sigma_B^2$
<i>St/B</i>	$2(r/2 - 1)$	$\sigma_e^2 + \sigma_{\mathbf{St} \times \mathbf{Su}}^2 + \mathbf{pn}\sigma_{\mathbf{St}}^2$
<i>Su/A</i>	$2(n/2 - 1)$	$\sigma_e^2 + \sigma_{\mathbf{St} \times \mathbf{Su}}^2 + qr\sigma_{\mathbf{Su}}^2$
<i>A × B/T</i>	1	$\sigma_e^2 + \sigma_{\mathbf{St} \times \mathbf{Su}}^2 + \mathbf{n}\sigma_{\mathbf{A} \times \mathbf{St}}^2 + r\sigma_{\mathbf{B} \times \mathbf{Su}}^2 + rn\sigma_{\mathbf{A} \times \mathbf{B}}^2$
<i>A × St/B</i>	$2(r/2 - 1)$	$\sigma_e^2 + \sigma_{\mathbf{St} \times \mathbf{Su}}^2 + n\sigma_{\mathbf{A} \times \mathbf{St}}^2$
<i>B × Su/A</i>	$2(n/2 - 1)$	$\sigma_e^2 + \sigma_{\mathbf{St} \times \mathbf{Su}}^2 + r\sigma_{\mathbf{B} \times \mathbf{Su}}^2$
<i>St/B × Su/A</i>	$4(n/2 - 1)(k/2 - 1)$	$\sigma_e^2 + \sigma_{\mathbf{St} \times \mathbf{Su}}^2$ ^c

a. *A* = Subject groups ($p = 2$), *B* = Stimulus sets ($q = 2$), *St* = Stimuli (r), *Su* = Subjects (n)

b. Variance components typed in bold letters are parts of the random – but not fixed – effects model.

c. σ_e^2 and $\sigma_{\mathbf{St} \times \mathbf{Su}}^2$ can not be estimated independently with only one observation per stimulus-subject combination.

$$F' = \frac{MS_{\mathbf{A} \times \mathbf{B}} + MS_{\mathbf{S} \mathbf{U}(\mathbf{A}) \times \mathbf{S} \mathbf{T}(\mathbf{B})}}{MS_{\mathbf{B} \times \mathbf{S} \mathbf{U}(\mathbf{A})} + MS_{\mathbf{A} + \mathbf{S} \mathbf{T}(\mathbf{B})}} \tag{4}$$

One straightforward way to avoid such statistical problems is to confound stimuli and the subjects factor. Under these circumstances, the fixed-effects model can be applied. In the present case, however, in order to confound both factors it is not necessary to draw a new random sample of stimuli for each subject (e.g., Keppel, 1976; Coleman, 1979; Richter & Seay, 1987). Instead, it is sufficient to randomly assign stimuli to conditions individually for each subject. That is, the distribution of the test statistic is not generated post hoc by means of permutations, like in the randomization test, but the permutation is actually performed individually for each subject.

7 Summary and Conclusions

I will divide the conclusions in two parts. First, I will discuss the methodological and statistical conclusions with respect to the treatment of stimuli in experimental and quasi-experimental studies. Second, I will discuss general implications of the present analysis for the question of what exactly a sound conceptualization of the randomness of random variables in psychological research could be.

7.1 Methodological Conclusions

The methodological conclusions for the treatment of stimuli in the three designs considered are straightforward:

1. If stimuli and treatment are confounded, there is no point in treating stimuli as random effects, because this approach usually will not have any positive effect on the internal validity of the study – at least if one does not have reasons to believe in the unlikely a-priori assumption that treatment and PCVs

are stochastically independent in the population. Internal validity should be sought by common (quasi-) experimental control techniques, like blocking and proper selection of material. Internal validity can not be reached by simply treating stimuli as random effects. However, the control of PCVs is possible to a higher degree with respect to stimuli than with respect to subjects.

2. If stimuli and treatment conditions are crossed – i.e., the design is "within-stimuli" – there is no sense in treating stimuli as random, because there is no reasonable notion of "chance fluctuations" with respect to stimuli. Therefore, stimulus variability has not to be controlled by inferential statistics in order to secure internal validity.
3. If stimuli and treatment conditions are counterbalanced, there is a symmetrical random partitioning of stimuli and subjects. Consequently, both, stimuli and subjects, should be treated as random effects, in order to account for any effects caused by the randomization procedure itself. To avoid technical problems with the handling of *Quasi-F* statistics, it is advised to confound stimuli and treatment conditions by randomizing the stimuli individually for each subject. In this case the fixed-effects model can be applied.

7.2 General Conclusions

The conceptual starting point of the present analysis was the idea of taking seriously the interpretation of parametric significance tests as approximations of randomization tests. Specifically, randomization procedures have been directly related to variance components in the fixed- and random-effects ANOVA. On the one hand, randomization tests reflect the logic and purpose of significance testing in experimental research in two major respects: (1) A statistical generalization to an underlying population is not intended, in contrast, all conclusions are principally sample related, and (2) the significance test protects from falsely accepting a causal hypothesis and is therefore supporting the internal validity of a study. On the other hand, ANOVA-procedures are widely used as a flexible mathematical tool for the analysis of experimental data, allowing for the treatment of different sources of variance as either fixed or random. The underlying statistical assumptions concerning the idea of statistical generalization to a population, however, should not be confused with the scientific aim of experimentation in psychology.

To state it somewhat differently, parametric statistics and randomization tests reflect two different facets of randomness. The first kind of randomness is realized by the randomization procedure in randomization tests. The randomization test protects either from errors of randomization caused by the random assignment itself, or from chance fluctuation not accounted for by the randomized sequence of treatment levels in within-subjects designs.

The second kind of randomness is expressed in the concept of random sampling from a population. In the present experimental context, this random sampling is given *eo ipso* with respect to some underlying *hypothetical population*, that is, it has not to be assumed that the sample is representative with respect to some underlying *actual population*. Based on these assumptions, what are the more general conclusions, that can be drawn from the present analysis? The first major conclusion concerns the status of inferential statistics in quasi-experiments or correlational analyses. It has been shown with respect to stimuli that the appeal to an underlying population – probably supported by the statistical assumptions of the ANOVA procedures – has the aim to protect the intended causal inference. This logic requires, however, that the underlying – hypothetical – stimulus populations do not differ with respect to the distributions of PCVs, that is, the validity of the *ceteris-paribus* condition has to be postulated on the level of – hypothetical –

populations. It has been argued that apart from the fact that PCVs – and their distributions in hypothetical populations – are principally better controllable with respect to stimuli than to subjects, this reasoning is also true for all other types of studies in which a causal inference is sought, but randomization is not possible. Curiously then, in quasi-experiments and correlational analyses, the reference to an underlying population becomes necessary not because an inference to these populations is sought but because it is necessary for the manifestation of the internal validity in the sample. As already indicated, the assumption that treatment and PCVs are stochastically independent in the – hypothetical – populations is crucial in this context. To explicate further, if PCVs and the hypothetical cause(s) are not stochastically independent in the population – i.e., PCVs are actually confounding variables (ACVs) – the actual probability of falsely rejecting the null-hypothesis (on the level of the scientific hypothesis) might be *higher* (if the confounding variables work in the direction of the causal hypothesis) or *lower* (if the confounding variables work against the causal hypothesis) than the nominal α -level.

Unfortunately, the assumption of stochastic independence is very difficult to prove empirically since most likely not all PCVs are known, to begin with. This would require the unreachable ideal of a "complete psychological theory". What one can do is at first to argue in favor of the validity of this assumption, based on substantial theoretical reasoning, and secondly to try to control as much PCVs as possible. If one could in fact match subjects with respect to *all* PCVs the validity of a causal inference would be *guaranteed*. Incidentally, this idea of a "perfect matching" quite exactly mirrors similar conceptions in philosophical theories of probabilistic causality and causal explanation, like the concept of "objectively homogenous reference classes" in Salmon's (1984) theory of causal explanation. Although one can never reach this "perfect matching", one definitely can *not* leave this problem to the statistical inference.

Alternatively, in regression or path analysis the idea of a "perfect matching" is reflected by the methodological requirement to include all causally relevant variables – either alone or in combination – in a path analysis in order to be able to interpret paths in a causal fashion (the so-called "closedness" condition). This inclusion of other causally relevant variables – which change or mediate the critical causal relation – is also to be found in philosophical conceptions of probabilistic causation usually referred to as the "screening off" of other conditions (e.g., Davis, 1988). Note, that the conclusions based on sequential testing of variables in this respect (i.e., whether they mediate or change the influence of the hypothetical cause on the dependent variable) are principally dubious since the mediating effects could take place in any possible *combination* of other variables. As a result, any evaluation of a variable in terms of a causal relation is essentially preliminary if not all relevant variables are included in the analysis *at the same time*.

The second major conclusion concerns the concept of external validity and its relation to both kinds of randomness. It has been shown that it is important to conceptually distinguish between variability that originates either from the randomization procedure (in case of between-subjects designs) or from chance fluctuations across points of measurement (in case of within-subjects designs) and manifest variability caused by subject/stimulus-treatment interactions. For stimuli all respective variability was argued to be manifest, since there is no reasonable notion of chance fluctuation with respect to stimuli. With respect to subjects, both sources of variance are statistically indistinguishable but should nevertheless be separated conceptually. Whereas variability originating from the process of randomization or chance fluctuations is a potential threat to the internal validity of a study, variability caused by subject/stimulus-treatment interaction is not. As a consequence, considering the variability caused by stimulus-treatment interactions does not raise the internal validity of a study. Rather differential effectiveness of the treatment

is an aspect of "the concept formerly known as external validity". This aspect has been named *aggregation validity* (AV) to set it apart from population or stimulus validity (cf. Hager & Westermann, 1983). AV does not require the notion of an underlying population and is therefore exclusively sample related – as is internal validity. Of course, the evaluation of AV might require replication of an experiment with different-possibly selected-samples of subjects. However, one should distinguish between the empirical *evaluation* of AV and the *concept* of AV, the latter clearly needing no appeal to a population. The term AV has been chosen to emphasize that its focus lies in the evaluation of the validity of aggregation across stimuli and subjects.

For methods of isolating subjects for which the treatment has differential effectiveness – thus violating the hypothesized causal law – and who therefore cause disordinal subject-treatment interactions, see Iseler (1996b). From the present perspective the major aim is to separate variability originating from the randomization process itself from variability which is caused by manifest subject-treatment interaction. Inasmuch as there is manifest differential effectiveness of the treatment – i.e., homogenous subgroups of subjects with different treatment effects can be identified – aggregation across these subjects is not a valid procedure and therefore threatens the aggregation validity of an experiment.

References

- [1] Bonge, D. R., Schuldt, W. J., & Harper, Y. Y. (1992). The experimenter-as-fixed-effects fallacy. *The Journal of Psychology*, *126*, 477-486.
- [2] Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente. [Theory and design of psychological experiments.]* Darmstadt: Steinkopff.
- [3] 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.
- [4] Cohen, J. (1976). Random means random. *Journal of Verbal Learning and Verbal Behavior*, *15*, 261-262.
- [5] Coleman, E. B. (1964). Generalizing to a language population. *Psychological Reports*, *14*, 219-226.
- [6] Coleman, E. B. (1979). Generalization effects vs. random effects: Is σ_{TL}^2 a source of Type 1 or Type 2 error? *Journal of Verbal Learning and Verbal Behavior*, *18*, 243-256.
- [7] Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation. Design and analysis issues for field settings.* Chicago: Rand McNally.
- [8] Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York: Wiley.
- [9] Davis, W. A. (1988). Probabilistic theories of causation. In J. H. Fetzer (Ed.), *Probability and causality. Essays in honor of Wesley C. Salmon* (pp. 133-160). Dordrecht: Reidel.
- [10] Edgington, E. S. (1995). *Randomization tests* (3rd ed.). New York: Dekker.
- [11] Erdfelder, E., & Bredenkamp, J. (1994). Hypothesenprüfung. [Hypothesis-testing.] In W. H. Tack & T. Herrmann (Eds.), *Enzyklopädie der Psychologie. Themenbereich B: Methodologie und Methoden. Serie I: Forschungsmethoden der Psychologie. Band 1: Methodologische Grundlagen der Psychologie.* (pp 604-648) Göttingen: Hogrefe.
- [12] Fontenelle, G. A., Phillips, A. P., & Lane, D. M. (1985). Generalizing across stimuli as well as subjects: A neglected aspect of external validity. *Journal of Applied Psychology*, *70*, 101-107.
- [13] Gadenne, V. (1984). *Theorie und Erfahrung in der psychologischen Forschung. [Theory and empirical knowledge in psychological research.]* Tübingen: Mohr.

- [14] Hager, W., & Westermann, R. (1983). Planung und Auswertung von Experimenten. [Design and evaluation of experiments.] In J. Bredenkamp & H. Feger (Eds.), *Hypothesenprüfung. Enzyklopädie der Psychologie, Serie Forschungsmethoden, Bd. 5*. (pp. 24-238). Göttingen : Hogrefe.
- [15] Hopkins, K. D. (1984). Generalizability theory and experimental design: Incongruity between analysis and inference. *American Educational Research Journal*, *21*, 703-712.
- [16] Iseler, A. (1996a). A paradoxical property of aggregate hypotheses referring to the order of medians. *MPR-Online*, *1(4)*. URL: <http://www.hsp.de/MPR/>.
- [17] Iseler, A. (1996b). Populationsverteilungen von Merkmalen und Geltungsbereiche individuenbezogener Aussagen als Gegenstand der Inferenzstatistik in psychologischen Untersuchungen. [Population-distributions of features and the range of validity of single-subject related statements as a topic of inferential statistics in psychological research]. Paper presented at the 40. Kongreß der Deutschen Gesellschaft für Psychologie: München. URL: <http://userpage.fu-berlin.de/iseler/>.
- [18] Kenny, D. A., & Smith, E. R. (1980). A note on the analysis of designs in which subjects receive each stimulus only once. *Journal of Experimental Social Psychology*, *16*, 497-507.
- [19] Keppel, G. (1976). Words as random variables. *Journal of Verbal Learning and Verbal Behavior*, *15*, 262-265.
- [20] Maxwell, S. E., & Bray, J. H. (1986). Robustness of the Quasi-F statistic to violations of sphericity. *Psychological Bulletin*, *99*, 416-421.
- [21] Richter, M. L., & Seay, M. B. (1987). ANOVA designs with subjects and stimuli as random effects: Applications to prototype effects on recognition memory. *Journal of Personality and Social Psychology*, *53*, 470-480.
- [22] Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.
- [23] Santa, J. L., Miller, J. J., & Shaw, M. L. (1979). Using *Quasi-F* to prevent alpha inflation due to stimulus variation. *Psychological Bulletin*, *86*, 37-46.
- [24] Scheffé, H. (1963). *The analysis of variance*. New York: Wiley.
- [25] Scheirer, C. J., & Geller, S. E. (1979). The analysis of random effects in modeling studies. *Child Development*, *50*, 752-757.0
- [26] Siemer, M. (1993). Aspekte einer vollständigeren Analyse der Unvollständigkeit psychologischer Hypothesen. [Aspects of a more complete analysis of the incompleteness of psychological hypotheses.] In L. Montada (Ed.), *Bericht über den 38. Kongreß der Deutschen Gesellschaft für Psychologie, Bd. 2* (pp. 728-734). Göttingen: Hogrefe.
- [27] Steyer, R. (1992). *Theorie kausaler Regressionsmodelle. [The theory of causal regression-models.]* Stuttgart: Fischer.
- [28] Steyer, R., Gabler, S., & Rucai, A. A. (1995). Individual causal effects, average causal effects, and unconfoundedness in regression models. In F. Faulbaum & W. Bandilla (Eds.), *SoftStat '95. Advances in statistical software*, *5* (pp. 203-210). Stuttgart: Lucius & Lucius.
- [29] Wickens, T. D., & Keppel, G. (1983). On the choice of design and test statistic in the analysis of experiments with sampled materials. *Journal of Verbal Learning and Verbal Behavior*, *22*, 296-309.