

# A common misapplication of statistical inference: nuisance parameter control with null-hypothesis rejection tests

Jona Sassenhagen      Phillip M. Alday

February 2016

## Abstract

Much experimental research on behavior and cognition rests on stimulus or subject selection where not all parameters can be fully controlled, even when attempting strict matching. For example, when contrasting patients to controls, factors such as intelligence or socioeconomic status are often correlated with patient status. Similarly, when presenting word stimuli, factors such as word frequency are often correlated with primary variables of interest. One procedure very commonly employed to control for such nuisance parameter effects is conducting inferential tests on confounding stimulus or subject characteristics. For example, if word length is not *significantly* different for two stimulus sets, they are considered as matched for word length. Such a test has high failure rates and is conceptually misguided. It stems from a common misunderstanding of statistical tests. We show inferential testing to be inappropriate both pragmatically and philosophically, present a survey showing its high prevalence, and briefly discuss an alternative in the form of regression including nuisance parameters.

## Introduction

Methods sections in many issues of *Brain & Language* and similar journals feature sentences such as

Animate and inanimate words chosen as stimulus materials did not differ in word frequency ( $p > 0.05$ ).

Controls and aphasics did not differ in age ( $p > 0.05$ ).

In the following, we discuss the inappropriateness of this practice. A common problem in brain and behavioral research, where the experimenter cannot freely determine every stimulus and participant parameter, is the control of confounding/nuisance parameters. This is especially common in studies of language.

Typically, word stimuli cannot be constructed out of whole cloth, but must be chosen from existing words (which differ in many parameters). Stimuli are processed by subjects in the context of a rich vocabulary; and subject populations have usually been exposed to very diverse environments and events in their acquisition of language. A similar problem exists, for example, when comparing controls to specific populations, such as bilingual individuals or slow readers. The basic question researchers are faced with is then to prevent reporting e.g. an effect of word length, or bilingualism, when the effect truly stems from differences in word frequency, or socioeconomic status, which may be correlated with the parameters of interest. A prevalent method we find in the literature, namely inferential testing of stimuli, fails to perform the necessary control.

## NHST and nuisance control

Often, researchers will attempt to demonstrate that stimuli or participants are selected so as to concentrate their differences on the parameter of interest, i.e. reduce confounds, by conducting null-hypothesis testing such as  $t$ -tests or ANOVA on the potentially confounding parameter in addition or even instead of showing descriptive statistics in the form of measures of location and scale. The underlying intuition is that these tests establish whether two conditions differ in a given parameter and serve as proof that the conditions are “equal” on that parameter. This is, in turn, based on the common, but wrong intuition that significance in null hypothesis significance testing establishes that a contrast shows a *meaningful effect*.

In practice, we find insignificant tests are used as a necessary (and often sufficient) condition for accepting a stimulus set as “controlled”. This approach fails on multiple levels.

- Philosophically, these tests are inferential tests being performed on closed populations, not random samples of larger populations. Statistical testing attempts to make inferences about the larger population based on randomly selected samples. Here, the “samples” are not taken randomly, and we are not interested in the population they are drawn from, but in the stimuli themselves. For example, in a study on the effects of animacy in language processing, we do not care whether the class of animate nouns in the language is on average more frequent than the class of inanimate nouns. Instead, we care whether the selection of animate nouns *in our stimuli* are on average more frequent than the selection of inanimate nouns *in our stimuli*. But inferential tests answer the former question, not the latter. They refer to the population of stimuli that will largely not be used, or the population of subjects that will not be investigated, in the study.
- Pragmatically, beyond being inappropriate, this procedure does not test a hypothesis of interest. This procedure tests the null hypothesis of “the populations that these stimuli were sampled from do not differ in this

feature”, but what we are actually interested in is “the differences in this feature between conditions is not responsible for any observed effects”. In other words, this procedure tests whether the conditions differ in a certain respect to a measurable degree, but not whether that difference actually has any meaningful influence on the result.

- Additionally, these tests carry all the usual problems of Null Hypothesis Significance Testing (cf. Cohen 1992), including its inability to accept the null hypothesis directly. This means that even if the conditions do not differ significantly, we cannot accept the hypothesis that they do not differ; we can only say that there is not enough evidence to exclude this hypothesis (which we are not actually interested in). In typical contexts (e.g. setting the Type I rate to the conventional 5% level), the power to reject the null hypothesis of no differences is low (Button et al. 2013) due to a small number of items, meaning that even comparatively large differences may be undetected, while in larger stimulus sets, even trivially small differences may be rejected. Especially with small samples (e.g., 10 subjects per group, or 20 items per condition), the probability of detecting moderate confound effects is thus low – even if there are substantial differences, tests will not reject the null hypothesis, and stimulus sets might be accepted as being balanced based on a test with a low probability of rejecting even moderately imbalanced samples of such a size.

In other words, these tests are incapable of actually informing us about the influence of potential confounds, but may give researchers a false sense of security. This inferential stage offers no benefit beyond examining the descriptive measures of location and scale (e.g. mean and variance) and determining if the stimuli groups are “similar enough”. For perceptual experiments, there may even be established discrimination thresholds below which the differences are considered indistinguishable. The preferred solution is directly examining to what extent these potential confounds have an influence on the results, such as by including these nuisance parameters in the statistical model. This is readily implemented via multiple regression, particularly “mixed-effect” approaches (Gelman and Hill 2006; Fox 2016).

## Randomization checks in clinical research

In the context of baseline differences between treatment and control groups in clinical trials, a similar debate has been waged (e.g. Senn 1994). The procedure is called a “randomization check” as it refers to checking if assignment of subjects to treatments has truly been performed randomly. Philosophically, this is somewhat less misguided, but has also been determined to be pragmatically pointless. In truly experimental research such as clinical trials, the effect of treatment is the variable of interest, and true randomization can be performed with regards to the multitude of other factors that might influence results. But in the case of non-medical, quasi-experimental research (i.e. research where full control is not

possible and thus confounds are unavoidable), stimuli or subjects are typically *known* to not have been selected randomly, but by specific criteria (e.g., animate vs. inanimate nouns, or patients with vs. without a particular lesion). That is, in the case of medical studies with randomization checks, experimental validity is achieved by selecting subjects from a given population and randomizing their assignment to treatment. In the case of studies in the brain and behavioral sciences, stimuli are constructed so as to differ on one parameter which we highly expect to be correlated with other parameters, e.g. word frequency and word length, and the concern of researchers is not if assignment was random (in fact, it is known to not have been random), but if stimuli differ systematically on variables expected to impact the dependent variable of interest. We are not aware of similar discussions in the psychological, linguistic or neurocognitive literature. Nonetheless, the clinical trial literature provides important considerations for experimental design choices e.g. by discussing the proper way of blocking and matching (Imai, King, and Stuart 2008), and can thus inform preparing stimulus sets or participant groups even for non-clinical experiments.

## Prevalence

We performed a literature survey of neurolinguistic studies to estimate the prevalence of inferential tests of nuisance parameters.

### Qualitative impressions

Instances of the error can be easily found not only in recent, but also in older publications, such as this example from the 1980s:

the two prime categories were equivalent in text frequency ([...] et al., 1971), and in length (both  $t$ 's  $< 1.1$ )

Here, the authors demonstrate in one sentence many of the fallacies underlying this procedure: both estimating a known quantity, and deducing equivalence (acceptance of the null) from a failed test (in this case, a test that leads the authors to accept the null hypothesis). To estimate how common the problem is in neurolinguistics, a high-quality neurolinguistic journal, *Brain & Language*, was investigated.

### Quantitative prevalence of the problem in recent issues of *Brain & Language*

In total, 86 articles were found where researchers reported known quantities in their stimulus/materials section, and 58 (**67%**) of these reported inferential

statistics of these known values. Of these, 47 (**81%**) “accepted” the null hypothesis (i.e., implicitly assumed that stimuli or subjects were matched following a nonsignificant test). We conclude that in a large fraction of those cases where researchers published in *B&L* are concerned about nuisance parameters of experimental stimuli, they conduct meaningless tests and misinterpret the results of these tests in a potentially misleading manner.

## Discussion and recommendation

In sum, NHST control of nuisance parameters is prevalent and meaningless, based on a flawed application of statistics to an irrelevant hypothesis. Luckily, proper nuisance control (of known and measurable variables) is not complex, although it can require more effort and computer time.

Researchers should still use descriptive statistics to demonstrate the success of balancing, but beyond that,  $p$  values from statistical tests on the stimulus properties offer no reliable, objective guideline. To directly and objectively estimate the influence of a set of stimuli on the dependent variables of interest, researchers should include stimulus properties in their statistical model for the data. For traditional  $t$ -tests, ANOVAs and regression models, this corresponds to using multiple regression with the stimuli properties as additional nuisance parameters (including continuous factors). In multiple regression, all parameters are jointly estimated and the total variance is allocated over all parameters depending on their independent impact. Thus, a condition effect estimated by a model also containing nuisance parameters corresponds to the effect of condition while accounting for nuisance parameter influence. Importantly, to prevent  $p$ -value fishing, the choice of selecting covariates to include must be made on principled grounds, and either a priori, or via unbiased model selection procedures.

Hierarchical/multilevel modeling (a.k.a mixed-effects modeling; see also Pinheiro and Bates 2000; Gelman and Hill 2006; Fox 2016) provides the necessary extension to the regression procedure for repeated-measures designs. Multilevel regression models have the additional advantage of accounting for the combined variance of subjects and items in one model (Clark 1973; Baayen, Davidson, and Bates 2008; Judd, Westfall, and Kenny 2012).

One problem in this context is that these stimulus confounds can be assumed to be correlated not only with one another and the dependent variables, but often also with the independent variables of interest (e.g., word frequency and word length correlate). This leads to model collinearity – the problem that models become hard to estimate and have biased variance estimates due to strongly correlated parameters. Popular software for mixed-effects models such as `lme4` automatically provides a summary of correlation between effects (Bates et al. 2015).

The main technique for dealing with collinearity is one that researchers traditionally already employ: attempting to balance stimulus/subject selection so that differences in nuisance parameters are minimised, e.g. via matching or blocking. That is, matching should generally still be performed in addition to multivariate estimation. However, often, they can not be entirely abolished, and significant correlations remain. Even in these cases, multiple regression is still the preferred solution if the present collinearity is dealt with appropriately.

Calculating such complex regression models will of course require more data, as power is lost with each additional parameter being estimated. We view this as a good thing because studies in the brain and behavioral sciences are chronically underpowered (Button et al. 2013).

Thus, our recommendations for the control of nuisance parameters are:

- attempt to control stimulus parameters to a reasonable degree
- use descriptive, but not inferential statistics to guide stimulus selection
- add confounding parameters as covariates into the final data analysis process
- use high-powered samples

Each step in this list is (hopefully) uncontroversial and helpful, unlike null-hypothesis testing of stimulus balance.

## Acknowledgements

We thank Sarah Tune for helpful discussion and Tal Linzen for bringing to our attention the randomization check literature. This work was supported in part by the German Research Foundation (BO 2471/3-2) and by the ERC grant 617891.

## References

- Baayen, R. H., D. J. Davidson, and D. M. Bates. 2008. “Mixed-Effects Modeling with Crossed Random Effects for Subjects and Items.” *Journal of Memory and Language* 59: 390–412.
- Bates, Douglas, Martin Maechler, Benjamin M. Bolker, and Steven Walker. 2015. “Fitting Linear Mixed-Effects Models Using lme4.” *ArXiv*, 1406.5823.
- Button, Katherine S, John P A Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma S J Robinson, and Marcus R Munafò. 2013. “Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience.” *Nat Rev Neurosci*, Apr.
- Clark, Herbert H. 1973. “The Language-as-Fixed-Effect Fallacy: A Critique of

- Language Statistics in Psychological Research.” *Journal of Verbal Learning and Verbal Behavior* 12: 335–59.
- Cohen, Jacob. 1992. “A Power Primer.” *Psychological Bulletin* 112 (1): 55–159.
- Fox, John. 2016. *Applied Regression Analysis and Generalized Linear Models*. 3rd ed. Thousand Oaks, CA: Sage.
- Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
- Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2008. “Misunderstandings Between Experimentalists and Observationalists About Causal Inference.” *Journal of the Royal Statistical Society A* 171, Part 2: 481–502.
- Judd, Charles M., Jacob Westfall, and David A. Kenny. 2012. “Treating Stimuli as a Random Factor in Social Psychology: A New and Comprehensive Solution to a Pervasive but Largely Ignored Problem.” *J Pers Soc Psychol* 103 (1): 54–69.
- Pinheiro, José, and Douglas Bates. 2000. *Mixed-Effects Models in S and S-PLUS*. Springer New York.
- Senn, Stephen. 1994. “Testing for Baseline Balance in Clinical Trials.” *Statistics in Medicine* 13: 1715–26.

## Methods

### Survey

The analysis was restricted to current volumes. For all articles published by *B&L* from 2011 to the 3rd issue of 2013, three raters (not blinded to the purpose of the experiment) investigated all published experimental papers (excluding reviews, simulation studies, editorials etc.). For each experiment reported in a study, the stimulus/materials sections were investigated for descriptive and inferential statistics derived from populations that were exhaustively sampled without error. If a descriptive and/or inferential statistic (such as mean and standard deviation) were reported, the study was coded as one where the researchers were interested in a known quantity, otherwise it was discarded. If an inferential statistic (such as a  $p$ -value) was reported, the study was coded as one where researchers answered that interest with an erroneous parameter estimate, otherwise as one where researchers did not commit the error. If a statement of the form that groups were thought equivalent regarding the parameter was made, such as claims that they were “matched”, “equal” or “did not differ”, and this statement was backed up by a  $p$ -value greater than 0.05, the study was coded as “accepting the null”. In cases of rater disagreement, the majority vote was registered. Representative statements from every study committing an error are available online at <https://github.com/jona-sassenhagen/statfail>.