A statistics lesson

Herbert H. Clark
Department of Psychology
Stanford University

Experiments on language before 1973

Typical experiment

1. 2, 3, or so treatments
2. 20 subjects
3. 20 items (typically words)

ANOVA

1. Subjects were treated as ***sampled***
2. Items were treated as ***fixed, not sampled***

Treatment effect was "significant" if …

ANOVA by subjects was significant

In 1972, I submitted a paper treating items as a random effect. The editor said I couldn't do that!

The "language-as-fixed-effect fallacy" (1973)

JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR 12, 335–359 (1973)

The Language-as-Fixed-Effect Fallacy:
A Critique of Language Statistics in Psychological Research

HERBERT H. CLARK[1]
Stanford University

Current investigators of words, sentences, and other language materials almost never provide statistical evidence that their findings generalize beyond the specific sample of language materials they have chosen. Nevertheless, these same investigators do not hesitate to conclude that their findings are true for language in general. In so doing, it is argued, they are committing the language-as-fixed-effect fallacy, which can lead to serious error. The problem is illustrated for one well-known series of studies in semantic memory. With the appropriate statistics these studies are shown to provide no reliable evidence for most of the main conclusions drawn from them. A review of other experiments in semantic memory shows that many of them are likewise suspect. It is demonstrated how this fallacy can be avoided by doing the right statistics, selecting the appropriate design, and sampling by systematic procedures, or, alternatively, by proceeding according to the so-called method of single cases.

In 1964, Edmund B. Coleman published an important methodological paper called "Generalizing to a Language Population" replicated if a different sample of language materials were used (p. 219)." Coleman then described available statistical procedures that

Argument in Clark 1973

Three possible F-ratios for treatment effects

F1	with ***subjects*** as random effect
F2	with ***items*** as random effect
F′	with ***both*** as random effects simultaneously

Recommended

F′	"quasi-F ratio"
min F′	F1*F2/(F1 + F2)

Not recommended

F1	alone
F2	alone
F1•F2	→ **F1•F2 criterion:** An effect is significant, if F1 and F2 are each significant

Initial reaction

People ***hated*** Clark 1973

- Favorite effects were ***no longer significant***
- *F′* and *min F′* were ***too much trouble to compute***

People looked for reasons to reject Clark 1973

- *F′* is ***too conservative***
- *min F′* is ***even more conservative***

Typical reactions

1. ***Reject argument*** in Clark 1973 outright
2. Try to ***circumvent*** Clark 1973
3. Use the criterion ***F1•F2*** (despite Clark's evidence)

Clark became a verb:

Student overheard telling advisor: "I've been Clarked."

Wike & Church, Cohen, Smith, Keppel, 1976

Comments on Clark's "The Language-as-Fixed-Effect Fallacy"

EDWARD L. WIKE AND JAMES D. CHURCH
The University of Kansas

Clark's arguments for treating language materials as random rather than fixed effects are examined, and the problems with random effects designs and approximate statistical tests (quasi-*F* ratios) are reviewed. In view of the difficulties with Clark's recommended

These authors rejected Clark 1973:

1. *F′* is ***not a legitimate*** F-ratio
2. Experimenters ***don't really sample*** items

Clark replied:

1. *F′* was shown (in Monte Carlo study) to "do well" except under special circumstances
2. Experimenters always ***describe their items as sampled*** even though they aren't.

Forster & Dickinson, 1976

More on the Language-as-Fixed-Effect Fallacy:
Monte Carlo Estimates of Error Rates for F_1, F_2, F', and *min F′*

K. I. FORSTER AND R. G. DICKINSON
Monash University

1. *Min F′* tends to be a very close estimate of *F′*
2. *F′* and *min F′* yield a true alpha level that is correct
3. *F1*, *F2*, and *F1•F2* yield inflated alpha levels

priate test on the basis of preliminary tests of item variance and subject-by-treatment variance.

Clark (1973) has drav … *min F′* test proposed by deficiencies in the sta… rs to be a very rigorous number of recently p… it may be that some which seek to establish… been insufficiently im- the way in which lin… rs of treating variation processed. The essen… ed effect, and hence do

F1•F2 criterion
F1 significant, p < .05
F2 significant, p < .05
Then *true* alpha = .10

Santa et al. 1979

Psychological Bulletin
1976, Vol. 83, No. 1, 37-46

Using Quasi *F* to Prevent Alpha Inflation Due to Stimulus Variation

John L. Santa, John J. Miller, and Marilyn L. Shaw
Rutgers—The State University

The nominal alpha level may be very inflated in much of the published literature where the conventional *F* test is used. This alpha inflation is often caused by ignoring stimulus variation or treating it as a fixed effect. The present article

In "a variety of realistic situations in which the data violate distribution and homogeneity of variance assumptions" "the *Quasi F has proved to be robust*."

It has long been traditional in psychology to treat subjects as a random effect in analysis of variance (ANOVA). The rationale is simple: For each experiment, we select from the population a small sample of subjects, but we want our results to generalize beyond this sample. Rarely are we interested in presenting our data as being pertinent only to the particular individuals studied. Both Clark (1973) and the exact distribution of the Quasi *F* statistic is unknown. However, when degrees of freedom are appropriately adjusted, the conventional *F* distribution can be used to approximate the Quasi *F* statistic (see Winer, 1971, p. 377).

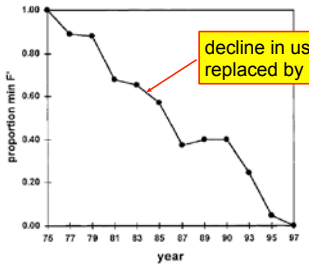
Clark applied the Quasi *F* analysis to a large body of semantic memory research and demonstrated that treatment effects can be

Others adopted Clark 1973 argument

1. Martindale 1978: The *therapist-as-fixed-effect fallacy* (in psychotherapy research)
2. Malgady et al. 1979: The *fixed-effect fallacy in educational psychological research*
3. Bonge et al 1992: The *experimenter-as-fixed-effect fallacy*
4. Judd et al. 2003: *Treating stimuli as a random factor* in social psychology
5. Monk 2004: The *product as fixed-effect fallacy*

But statistical practices gradually eroded

Raaijmakers et al. 1999



decline in use of *min F'* replaced by *F1*•*F2*

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

11

Raaijmakers et al. 1999

How to Deal with "The Language-as-Fixed-Effect Fallacy": Common Misconceptions and Alternative Solutions

Jeroen G. W. Raaijmakers
University of Amsterdam, Amsterdam, The Netherlands
and
Joseph M. C. Schrijnemakers and Frans Gremmen
University of Nijmegen, Nijmegen, The Netherlands

Although Clark's (1973) critique of statistical procedures in language and memory studies (the

Many investigators assume that *F1*•*F2* is justified. "Such a procedure is, however, unfounded and not in accordance with the recommendations of Clark (1973)."

Key Words: design; *minF'*; language; fixed effect; random effect.

Suppose that in a primed lexical decision variable (the individual word pairs) is a fixed

12

Raaijmakers et al. 1999

How to Deal with "The Language-as-Fixed-Effect Fallacy": Common Misconceptions and Alternative Solutions

Jeroen G. W. Raaijmakers
University of Amsterdam, Amsterdam, The Netherlands

And yet, Raijmakers et al proposed using ***F1*** for certain experimental designs.

Clark 1973 showed that F1 in these designs had ***high type I error rates***.

Baayen et al 2008 showed high type I error rates in Monte Carlo study

Key Words: design; *minF'*; language; fixed effect; random effect.

Suppose that in a primed lexical decision variable (the individual word pairs) is a fixed

13

In 2008, F' is *finally* replaced

Harald Baayen argued: ANOVAs should be replaced by "linear mixed effect" models

Baayen et al. 2008

ELSEVIER
Available online at www.sciencedirect.com
ScienceDirect
Journal of Memory and Language 59 (2008) 390–412
Journal of Memory and Language
www.elsevier.com/locate/jml

In Monte Carlo studies (like Forster 1976):

1. *Min F'* tends to be a very close estimate of *F'*
2. *F'* and *min F'* yield a true alpha level that is correct
3. *F1, F2,* and *F1*•*F2* yield inflated alpha levels

F1•F2 criterion
F1 significant, p < .05
F2 significant, p < .05
Then *true* alpha = .10

Abstract

This paper provides an

© 2007 Elsevier Inc. All

Keywords: Mixed-effects models; Crossed random effects; Quasi-F; By-item; By-subject

15

Barr, Levy, Scheepers, Tily, 2013

Journal of Memory and Language 68 (2013) 255–278
Contents lists available at SciVerse ScienceDirect
Journal of Memory and Language
journal homepage: www.elsevier.com/locate/jml
ELSEVIER

- Expansion on Baayen's "linear mixed-effects" models
- Monte Carlo simulations with similar findings

ARTICLE INFO

Article history:
Received 5 August 2011
revision received 30 October 2012

Keywords:
Linear mixed-effects models
Generalization
Statistics
Monte Carlo simulation

ABSTRACT

Linear mixed-effects models (LMEMs) have become increasingly prominent in psycholinguistics and related areas. However, many researchers do not seem to appreciate how random effects structures affect the generalizability of an analysis. Here, we argue that researchers using LMEMs for confirmatory hypothesis testing should minimally adhere to the standards that have been in place for many decades. Through theoretical arguments and Monte Carlo simulation, we show that LMEMs generalize best when they include the maximal random effects structure justified by the design. The generalization performance of LMEMs including data-driven random effects structures strongly depends upon modeling criteria and sample size, yielding reasonable results on moderately-sized samples when conservative criteria are used, but with little or no power advantage over maximal models. Finally, random-intercepts-only LMEMs used on within-subjects and/or within-items data from populations where subjects and/or items vary in their sensitivity to experimental manipulations always generalize worse than separate F_1 and F_2 tests, and in many cases, even worse than F_1 alone. Maximal LMEMs should be the 'gold standard' for confirmatory hypothesis testing in psycholinguistics and beyond.

16

What statistics lesson have we learned?

All experimenters maximize benefit/cost ratio

Benefits to self
- I discover an important new phenomenon
- I become known for the discovery

Benefits to science
- An important new phenomenon has been discovered
- It becomes known

Costs to self
- Type 2 errors

Costs to science
- Type 1 errors

The statistics lesson
1. Experimenters tend to weigh type 2 errors more than type 1 errors
2. Science tends to weigh type 1 errors more than type 2 errors

18

Thanks!