Editorial

Beyond the Significance Test Ritual

What Is There?

Peter Sedlmeier (Guest Editor)

Chemnitz University of Technology, Germany

The mindless use of null-hypothesis significance testing – the significance test ritual (e.g., Salsburg, 1985) – has long been criticized. The main component of the ritual can be characterized as follows: Once you have collected your data, try to refute your null hypothesis (e.g., no mean difference, zero correlation, etc.) in an automatized manner. Often the ritual is complemented by the "star procedure": If $p < .05$, assign one star to your results (*), if $p < .01$ give two stars (**), and if $p < .001$ you have earned yourself three stars (***). If you have obtained at least one star, the ritual has been successfully performed; if not, your results are not worth much. The stars, or the corresponding numerical values, have been door-openers to prestigious psychology journals and, therefore, the ritual has received strong reinforcement.

The ritual does not have a firm theoretical grounding; it seems to have arisen as a badly understood hybrid mixture of the approaches of Ronald A. Fisher, Jerzy Neyman, Egon S. Pearson, and (at least in some variations of the ritual) Thomas Bayes (see Acree, 1979; Gigerenzer & Murray, 1987; Spielman, 1974). For quite some time, there has been controversy over its usefulness. The debates arising from this controversy, however, have not been limited to discussions about the mindless procedure as sketched above, but have expanded to include the issues of experimental design and sampling procedures, assumptions about the size of population effects (leading to the specification of an alternative hypothesis), deliberations about statistical power *before* the data are collected, and decisions about Type I and Type II errors. There have been several such debates and the controversy is ongoing (for a summary see Balluerka, Gómez, & Hidalgo, 2005; Nickerson, 2000; Sedlmeier, 1999, Appendix C). Although there have been voices that argue for a ban on significance testing (e.g., Hunter, 1997), authors usually conclude that significance tests, if conducted properly, probably have some value (or at least do no harm) but should be complemented (or replaced) by other more informative ways of analyzing data (e.g., Abelson, 1995; Cohen, 1994; Howard, Maxwell, & Fleming, 2000; Loftus, 1993; Nickerson, 2000; Sedlmeier, 1996; Wilkinson & Task Force on Statistical Inference, 1999).

Alternative data-analysis techniques have been well-known among methodologists for decades but this knowledge, mainly collected in methods journals, seems to have had little impact on the practice of researchers to date. I see two main reasons for this unsatisfactory state of affairs. First, it appears that there is still a fair amount of misunderstanding about what the results of significance tests really mean (e.g., Gordon, 2001; Haller & Krauss, 2002; Mittag & Thompson, 2000; Monterde-i-Bort, Pascual Llobell, & Frias-Navarro, 2008). Second, although alternatives have been briefly mentioned in widely received summary articles (such as Wilkinson & Task Force on Statistical Inference, 1999), they have rarely been presented in a nontechnical and detailed manner to a nonspecialized audience. Thus, researchers might, in principle, be willing to change how they analyze data but the effort needed to learn about alternative methods might just be regarded as too great.

The main aim of this special issue is to introduce a collection of these alternative data-analysis methods in a nontechnical way, described by experts in the field. Before introducing the contents of the special issue, I will briefly outline the ideal state of affairs in inference statistics and discuss the difference between mindless and mindful significance testing.

The Ideal State of Affairs

Significance tests help us make inferences about population parameters, such as determining if population means differ or if two variables are really correlated. We use them to provide us with some information about what we do not have direct access to. What would be the ideal state of affairs when making an inference about population pa-

Zeitschrift für Psychologie / Journal of Psychology 2009; Vol. 217(1):1–5
DOI 10.1027/0044-3409.217.1.1

rameters? Obviously, the maximum possible knowledge would be to know the exact parameter values. This is a somewhat sobering thought, because what researchers and recipients of research results often really want to know is what their results *mean* – for instance, whether the difference between the means or the amount of correlation is practically significant or "really meaningful." Although the results of significance tests per se are sometimes interpreted as if they could answer questions of meaningfulness of results, of course they cannot. It is up to the judgment of the researcher or of the research community to interpret a given effect. This kind of judgment can only sensibly be made if it is grounded in knowledge of the subject matter.

What remains is the attempt to get as close as possible to the ideal state of affairs with statistical methods: Making estimates of population parameters as precisely as possible or at least having some information about how precise these estimates are. Often, however, simple estimates for population parameters will not do, because of the complexity of the research question. Especially in these cases, but also in simpler analyses, graphs can help us understand empirical data and reach tentative conclusions about the corresponding population values. Even more important and, in fact, indispensable for making sound inferences, are theoretical deliberations *before* a study is run. Such attempts to increase precision and detail will be discussed below, but first, let us have a look at how significance testing itself can be made more meaningful.

Mindful Significance Testing

R.A. Fisher, who arguably was the most influential figure in introducing significance tests to methodologically unsophisticated researchers, would certainly not have agreed with the significance test ritual as described above. Fisher regarded significance testing as a task that requires considerable methodological skill and profound knowledge about the subject matter, and he also spoke against giving single results too much weight (e.g., Fisher, 1929). His tenet that the null hypothesis can only be refuted and not accepted has become part of the significance test ritual, although J. Neyman and E.S. Pearson showed that significance testing (they called the procedure "hypothesis test") can be symmetrical (the null hypothesis can, in principle, also be accepted) by including an alternative hypothesis (e.g., Neyman, 1950). In their approach, before conducting an experiment, it is absolutely necessary to deliberate extensively about the possible effects of Type I and Type II errors (deciding against the null hypothesis if it is true or deciding against the alternative hypothesis if that is true), which are the basis for choosing the right sample size. The decision analysis contained in the Neyman-Pearson approach is rarely needed in psychological research and this might be one reason why the approach is not so widely used (see Hager, 2005). However, the hybrid approach that is more or less established in psychological research can also be conducted in a mindful way that, for instance, allows one to accept the null hypothesis (e.g., Aaron & Aaron, 2003; Rosenthal & Rosnow, 1991; Sedlmeier & Renkewitz, 2007).[1]

There have been several other suggestions for using significance tests in a more mindful way (e.g., Jones & Tukey, 2000; Killeen, 2005; Wainer, 1999; Westermann, 2000). One of the most useful extensions of the standard approach is the use of alternative hypotheses that are as specific as possible. One prominent example is contrast analysis (Rosenthal, Rosnow, & Rubin, 2000). However, even if significance tests are performed in a mindful way, the result, which can be seen as a yes/no answer to the question "Can the deviation from the expected value (specified by the null hypothesis) still be interpreted as being the result of chance?" is of limited usefulness. In most cases, this piece of information is far from the ideal, that is, exact knowledge about the population parameters in question.

Beyond the Ritual: Higher Precision, More Detail

Effect Sizes and Confidence Intervals

How can we get closer to this ideal? Two ways have been prominently proposed: Confidence intervals instead of *p*-values, and effect sizes. Rosnow and Rosenthal (this issue) will introduce the different varieties of effect sizes, including the calculation of effect sizes for contrast analyses, and simple confidence intervals for effect sizes. Cumming and Fidler (this issue) will then provide a comprehensive tutorial on confidence intervals, including a discussion of more sophisticated confidence intervals for effect sizes and a comparison of different ways to calculate them. There have been many studies on misunderstandings of significance test results but there is still little known about how well confidence intervals are understood by researchers and students. Fidler and Loftus (this issue) report evidence regarding the understanding of confidence intervals that may be seen as grounds for some optimism.

1 One central problem that prohibits mindful significance testing seems to be a common misunderstanding about what sampling distributions are and how sample size influences the shape of sampling distributions, and, therefore, statistical power (e.g., Oakes, 1986; Sedlmeier, 2006). Insight into the nature of sampling distributions might be fostered by performing computer simulations or using programs that show the impact of sample size on power (see Faul, Erdfelder, Lang, & Buchner, 2007; Sedlmeier, 2007).

Illustrations and Graphs

When one compares the relative prevalence of significance tests versus graphs in the "hard" (natural) sciences (e.g., chemistry, physics, and biology) with that in "soft" sciences (e.g., psychology, economics, and sociology) one finds that the hard sciences rely much more on graphs to communicate research results, whereas the soft sciences still heavily rely on significance tests (Smith, Best, Stubbs, Archibald, & Roberson-Nay, 2002). Interestingly, Smith et al. found the same situation in psychology: More graphs (and less significance tests) in journals that reported on results in "hard" areas of psychology (e.g., *Behavioral Neuroscience* and the *Journal of Experimental Psychology: Animal Behavior Processes*) than in journals that cover "softer" areas (e.g., the *Journal of Educational Psychology* and the *Journal of Clinical Psychology*). Does this mean that if data are softer, that is, if it is more difficult to find convincing results, graphs are not suitable for showing these results but significance tests are? I don't think so. Surely, more often than not, graphs are "better than a thousand *p*-values" (Loftus, 1993) for drawing conclusions from one's data, especially if they contain some ambiguity; and there is no better way to find details in the data than by using graphs or semigraphical displays (Tukey, 1977). There already exists a substantial collection of methods for visualizing data (e.g., Cleveland, 1994; Tufte, 1983; Wainer, 2005) that is still much underused in the analysis and communication of psychological research results. Kwan, Lu and Friendly (this issue) add one more valuable graphical tool that is well suited to the analysis of multivariate procedures, such as confirmatory factor analysis and goes well beyond the information one can get from significance test results.

More Precise Predictions

Being trained in the use of simple significance tests, such as *t*-test and analysis of variance, apparently has the potential to prompt researchers to develop simple theories, as well (Gigerenzer, 1991, 1998; see also Loftus, 1996): The world begins to look like a collection of m × *n* designs that can be easily examined by applying significance testing. In some cases, such as treatment-effectiveness studies, it might not be possible or even very necessary to derive more precise predictions, but most areas in psychology would certainly benefit from more precise theories (even if there is no significance test that fits). This increased precision can be accomplished with the use of mathematical models and computer simulations. Marewski and Ohlsson (this issue) give many details on how this can be achieved.

What Else is There Beyond the Ritual?

This special issue takes a look beyond the significance test ritual but, of course, it omits much more than it covers. There is, for instance, no treatment of special approaches to significance testing such as sequential methods (Wald, 1947) or randomization techniques (e.g., Efron & Tibshirani, 1993; Edgington, 1995); nor is meta-analysis covered (see, however, the 2007 issue of this journal *Zeitschrift für Psychologie/Journal of Psychology, 215*[2]). We also have completely omitted the "other world of inference statistics," Bayesian statistics, which was introduced to a wide audience of psychologists in 1963 by Ward Edwards and colleagues (Edwards, Lindman, & Savage, 1963) but still plays only a marginal role.[2]

All kinds of inference statistics rely on preconditions and assumptions that have to be at least partly fulfilled to allow a sound interpretation of the results (e.g., Hager, 2000; Haig, 2005; Westermann, 2000): We do not have much to say about these important issues either, because of space limitations. In addition, there are single-case designs (e.g., Kazdin, 1982), and qualitative methods (e.g., Smith, 2003) that can reach well beyond what significance test results can tell us. The long list of omissions should make clear that there is much to be found beyond the significance test ritual. There is, however, no single best solution in data analysis and there is definitely no ritual that could and should replace the significance test ritual. Experts in the field should continue to try to reach consensus through thoughtful consideration of the best methods available. This special issue makes some suggestions that may lead us in that direction.

Acknowledgments

Thanks to Friederike Brockhaus, Juliane Kämpfe, Thomas Schäfer, Anita Todd, and Isabell Winkler for their helpful comments.

References

Abelson, R.P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.

Acree, M.C. (1979). Theories of statistical inference in psychological research: A historico-critical study. *Dissertation Abstracts International*, *39*, 5073B. (UMI No. 7907000).

Aron, A., & Aron, E.N. (2003). *Statistics for psychology* (3rd ed). Upper Saddle River: Prentice Hall.

Balluerka, N., Gómez, J., & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology, 1*, 55–70.

2 Why this is so, how Bayesian statistics works, and how psychologists can benefit from using it might be worth another special issue.

Cleveland, W.S. (1994). *The elements of graphing data*. Summit, NJ: Hobart Press.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Edgington, E.S. (1995). *Randomization tests.* (3rd ed). New York: Marcel Dekker.

Edwards, W., Lindman, H., & Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap.* New York: Chapman and Hall.

Faul, F., Erdfelder, E., Lang, A-G., & Buchner, A. (2007). G*Power3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Fisher, R.A. (1929). The statistical method in psychical research. *Proceedings of the Society for Psychical Research*, *39*, 185–189.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, *98*, 254–267.

Gigerenzer, G. (1998). Surrogates for theories. *Theory and Psychology*, *8*, 195–204.

Gigerenzer, G., & Murray, D. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.

Gordon, H.R.D. (2001). American Vocational Education Research Association members' perceptions of statistical significance tests and other statistical controversies. *Journal of Vocational Educational Research, 26*(2), 1–18.

Hager, W. (2000). About some misconceptions and the discontent with statistical tests in psychology. *Methods of Psychological Research Online*, *5*, 1–31. Retrieved July 2008 from Internet: http://www.mpr-online.de

Hager, W. (2005). Vorgehensweisen in der deutschsprachigen psychologischen Forschung. Eine Analyse empirischer Arbeiten der Jahre 2001 und 2002. [German psychological research: An analysis of empirical studies from 2001 and 2002]. *Psychologische Rundschau*, *56*, 191–200.

Haig, B.D. (2005). An abductive theory of scientific method. *Psychological Methods*, *10*, 371–388.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance. A problem students share with their teachers? *Methods of Psychological Research Online*, *17*. Retrieved July 2008 from http://www.mpr-online.de

Howard, G.S., Maxwell, S.C., & Fleming, K.J. (2000). The proof of the pudding: Illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods*, *5*, 315–332.

Hunter, J.E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*, 3–7.

Jones, L.V., & Tukey, J.W. (2000). A sensible formulation of the significance test. *Psychological Methods*, *5*, 411–414.

Kazdin, A.E. (1982). Single-case research designs: *Methods for clinical and applied settings.* New York: Oxford University Press.

Killeen, P.R. (2005). An alternative to null hypothesis significance tests. *Psychological Science, 16*, 345–353.

Loftus, G.R. (1993). A picture is worth a thousand *p* values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments, and Computers*, *25*, 250–256.

Loftus, G.R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161–171.

Mittag, K.C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29*(4), 14–20.

Monterde-i-Bort, H., Pascual Llobell, J., & Frias-Navarro, D. (2008). *Uses and abuses of statistical significance tests and other statistical resources: A comparative study*. Unpublished manuscript, University of Valencia.

Neyman, J. (1950). *First course in probability and statistics*. New York: Henry Holt.

Nickerson, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* Chichester, UK: Wiley.

Rosenthal, R., & Rosnow, R.L. (1991). *Essentials of behavioral research: Methods and data analysis*. (2nd ed.). New York: McGraw-Hill.

Rosenthal, R., Rosnow, R.L., & Rubin, D.B. (2000). *Contrasts and effect sizes in behavioral research*. New York: Cambridge University Press.

Salsburg, D.S. (1985). The religion of statistics as practiced in medical journals. *The American Statistician*, *39*, 220–223.

Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. [Beyond the significance test ritual: Extensions and alternatives]. *Methods of Psychological Research Online, 1.* Retrieved July 2008 from http://www.mpr-online.de

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Mahwah, NJ: Erlbaum.

Sedlmeier, P. (2006). Intuitive judgments about sample size. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 53–71). Cambridge: Cambridge University Press.

Sedlmeier, P. (2007). Statistical reasoning: Valid intuitions put to use. In M. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 389–419). New York: Erlbaum.

Sedlmeier, P., & Renkewitz, F. (2007). *Forschungsmethoden und Statistik in der Psychologie* [Research methods and statistics in psychology]. Munich: Pearson Education.

Smith, J.A. (Ed.). (2003). *Qualitative psychology: A practical guide to research methods*. London: Sage.

Smith, L.D., Best, L.S., Stubbs, D.A., Archibald, A.B., & Roberson-Nay, R. (2002). Constructing knowledge: The role of graphs and tables in hard and soft psychology. *American Psychologist*, *57*, 749–761.

Spielman, S. (1974). The logic of tests of significance. *Philosophy of Science*, *41*, 211–225.

Tufte, E.R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Tukey, J.W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods, 4*, 212–213.

Wainer, H. (2005). *Graphic discovery: A trout in the milk and other visual adventures*. Princeton, NJ: Princeton University Press.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik*. [Philosophy of science and experimental methodology]. Göttingen: Hogrefe.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Peter Sedlmeier

Department of Psychology
Chemnitz University of Technology
D-09107 Chemnitz
Germany
Tel. +49 371 5-313-6431
Fax +49 371 5-312-7419
E-mail peter.sedlmeier@phil.tu-chemnitz.de