

## COMMENTARY

### What Constitutes Evidence in Science Education Research?

In the wake of an increasing political commitment to evidence-based decision making and evidence-based educational reform that emerged with the No Child Left Behind effort, the question of what counts as evidence has become increasingly important in the field of science education. In current public discussions, academics, politicians, and other stakeholders tend to privilege experimental studies and studies using statistics and large sample sizes. However, some science education studies use a lot of statistics and large sample sizes (e.g., Bodzin, 2011) and yet, as I suggest in this text, are flawed and do not provide (sound) evidence in favor of some treatment or claim. Leaving aside the assertion and consensus of researchers across the quantitative/qualitative spectrum (e.g., the collection of chapters in Ercikan and Roth, 2009), we must ask whether all studies that appear to provide “quantitative” support for a particular effect do in fact provide *substantial* or strong evidence. As an anonymous reviewer of this contribution has pointed out, the question in its title really has two dimensions: (a) What constitutes *valid* evidence and (b) what are the limits of the claims that can be constructed when the evidentiary chain from premises to results is perfectly constructed. Both are important in constructing explanations for phenomena of interest to scientists generally and to science educators in particular. I begin by discussing the two issues in the context of the logic of scientific inquiry and statistical inference and then exemplify the issues as these play out in one recent article published in the pages of this journal (Bodzin, 2011). To further concretize my discussion, I also sketch two re-analyses concerned with the weight of the evidence provided by (a) 10 studies of paranormal psychological phenomena (*psi*) and (b) 855 studies in experimental psychology.

#### Validity and the Logic of Scientific and Statistical Explanation

First, in the logic of science, all explanatory schemas – including those of historical, historical-developmental, or interpretive nature – can be expressed in the following way (e.g., Stegmüller, 1974). Some observed event  $E$  (i.e., the evidence) is related to the statements about antecedent conditions and general laws or law-like regularities; together these constitute the premises of the argument made in the research article. The conditions for an explanation to be valid include: (a) the argument that leads from a hypothesized regularity or law to observation has to be *correct*; (b) there has to be *at least one general law* or law-like regularity; (c) the hypothesized law/regularity has to include *empirical content*; and (d) the statements that constitute the law have to be *true* (based on basic logic, no valid inferences can be made otherwise). In the logic of experimental research, explanations may be of two kind: (a) given the same set of antecedent conditions, a first hypothesized law would lead to observed event  $E_1$  whereas a second hypothesized law leads to event  $E_2$ ; or (b) given the same law or law-like regularity, the antecedent conditions would lead to observed  $E_1$  whereas a second set of antecedent conditions would lead to  $E_2$ . Frequency-based statistics are used to establish the probability for an event  $E$  to be observed  $p(E|H_0)$  given the null hypothesis.<sup>1</sup> This probability gives only

---

<sup>1</sup> More technically expressed, the  $p$  value a study reports is the probability for a certain effect to occur given the null hypothesis. The probabilities are given by the appropriate distribution,



There are many other reasons why a valid inference nevertheless is problematic. It is up to the researcher to exhibit and discuss the strength of a study, the validity of its evidence, and which audiences will draw what kind of benefit from the study results.

### **Validity of Evidence in Science Education: An Example**

To illustrate how science educators might want to think about the strength of evidence they produce, I provide an exemplary look on a recent study in earth science education (Bodzin, 2011). Bodin suggests that the purpose of the study was to investigate the “extent [that] a [geospatial information technology (GIT)]-supported curriculum could help students at all ability levels . . . to understand [land use change (LUC)] concepts and enhance the spatial skills involved with aerial and RS imagery interpretation” (Bodzin, 2011, p. 293). That is, an explicit claim is made about a causal relationship between a curriculum and learning. Following the logic of argumentation outlined above, therefore, the author makes a claim that a particular antecedent (the GIT-supported curriculum) brings about a difference in achievement when the students are observed (tested) before and after the treatment. The study uses a simple pre-test (observation  $O_1$ ) / treatment ( $X$ ) / post-test ( $O_2$ ) design, which has the structure

$$\begin{array}{ccc} O_1 & X & O_2 \\ \hline \end{array}$$

(Cook & Campbell, 1979). In this case, the authors of the standard reference book on quasi-experimental design suggest “we should usually not expect hard-headed causal inferences from the simple before-after design when it is used by itself” (p. 103).

Although the authors suggest that such a design may produce hypotheses worthy of further exploration, they express the hope “that persons considering the use of this design will hesitate before resorting to it” (p. 103). This is so because the difference in test scores ( $O_2 - O_1$ ) could be due to maturation or other events in the life of the students (e.g., they learn certain mathematical concepts or concepts in logic). Because Bodzin does not rule out other reasonable alternatives, the design provides weak (little) to no evidence for a treatment effect because there are many other possible causes that could have brought about the differences in achievement between the two observations – even though the statistical tests are significant and even if the effect sizes were large.

If we accept for the moment that the study is exploratory, we may ask ourselves whether its evidence has any strength that warrants further study. We then have to choose the form of analysis. Traditionally, there is no question: the statistics would be one based on frequency distributions (e.g., Students  $t$ ). Within this frequency-based perspective, the evidence provided by Bodzin’s study is not strong even though it might appear as such. A first problem with the results is that the reported means are not independent because each overall means reported in each of Bodzin’s Tables 2, 3, and 4 really is a weighted mean derived from the other pieces of information already available. That is, it is as if the author reported that three individuals had \$2, \$3, and \$4, respectively, *and* also reported that they owned \$9 together or that the mean amount was \$3. The additional information is redundant rather than *additional* evidence; but reporting the redundant information makes it look like there is additional evidence. Statisticians tend to deal with this issue by lowering the degrees of freedom and thereby eliminating redundancy. The study therefore violates some basic assumptions for statistical inference that would be part of the second type of validity. Moreover, the overall means in his Table 2 can be calculated from the scales reported in Tables 3 and 4. To draw any useful conclusions from the  $p$  values,

however, the tests need to be independent. As presented, the study overestimates the evidence in favor of the treatment.

A second major problem is the number of  $t$  tests conducted: a total of  $N = 24$ , which, given the content of bullet 1 above, tremendously increases the possibility of a type I error. That is, the experiment-wise error rate that there is a false positive actually exceeds 1 ( $24 \cdot \alpha = 24 \cdot 0.05 = 1.2$ ) and, therefore, would be set to  $p = 1$  in statistical packages such as SAS.<sup>4</sup> To hold the *experiment-wise* error rate at  $\alpha = 0.05$ , tests could be adjusted using what is known as the Bonferroni procedure (or one of its alternatives).<sup>5</sup> In this procedure, every test in an ensemble of  $N$  tests is conducted at a revised  $\alpha$ -level of  $\alpha_{\text{new}} = \alpha/N$  so that the total, experiment-wise error still is less than  $\alpha = 0.05$ . That is, instead of a cut-off at  $p < 0.05$ ,  $p < 0.01$ , . . . the new cut-offs for rejecting the null hypothesis would be at  $p < 0.0021$ ,  $p < 0.00042$ , . . . and so on. Again, the reported tests are strongly biased in favor of the reported effects because these are conducted at error probabilities 24 times higher than acceptable. Another option would have been to use a MANOVA, that is, a test with multiple (“M-”) dependent measures tested simultaneously (“ANOVA”). Only when this test suggests a significant difference would more conservative, adjusted  $t$ -tests be warranted.

Even if these problems did not exist, further caution would be required because frequency-based statistics have some fundamental problems, even flaws. In a recent article of the *Journal of Personality and Social Psychology*, a group of authors reanalyzes the results of a set of experimental studies to respond to their rhetorical question “Why psychologists must change the way they analyze their data?” (Wagemakers, Wetzels, Borsboom, & van der Maas, 2011). This study was designed as a critique of a series of studies on the psychological phenomenon of *psi* all conducted by the same researcher (Bem, 2011). Here, the “term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms” (p. 407). It is a descriptive term that includes, among others, telepathy, clairvoyance, precognition, and premonition. The subject is a controversial one that – recognized as such by the study’s author – most psychologists right out reject even though there are significant parts of the general population believing in parapsychological phenomena. The series of studies has fulfilled all the criteria required by the logic of science for valid inference.

The study suggests that there is overwhelming, cumulative evidence for the existence of certain *psi*-related phenomena. However, the critique shows that even though there are nine (of 10) experimental studies conducted by Bem with statistically reliable results in favor of rejecting of the null hypothesis ( $H_0$ ) – i.e.,  $H_1 =$  there is no *psi* [precognition] – much of the evidence is only “anecdotal” in favor of either the null hypothesis (there is no *psi*) or its alternative (there is *psi*). To provide evidence for their counter claim, Wagemakers, Wetzels et al. (2011) use a simple Bayesian test that uses the unbiased prior possibilities but is not biased against the null hypothesis as is the frequency based statistics Bem, following standard procedure, used in his study.<sup>6</sup> Another investigation

<sup>4</sup> [http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug\\_multtest\\_sect014.htm](http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_multtest_sect014.htm)

<sup>5</sup> The procedure is sometimes critiqued for being too conservative.

<sup>6</sup> A calculator for this statistics is available at <http://pcl.missouri.edu/bayesfactor>. The website also provides access to relevant articles.

reanalyzes 855 studies in experimental psychology and suggests that 70% of the studies with  $0.01 < p < 0.05$  (i.e., a total of 132 studies) provide no more than anecdotal evidence for the effect of interest (Wetzels et al., 2011). That is, in a field that prides itself for the strength of methodological approaches, a large number of studies that appear to support the alternative to the null hypothesis of no (treatment) effect actually provide evidence that is at best anecdotal.

Bayesian statistics have been proposed as a way of overcoming many of the inherent problems with frequency-based statistics not in the least because it allows researchers to quantify prior knowledge (e.g., Gurrin, Kurinczuk, & Burton, 2000). Bayesian statistics are of such a nature that they can be used to provide direct and explicit answers to questions that are usually posed by practitioners. This is so because Bayesian statistics asks what the probability  $p(H_0|E)$  for the null hypothesis  $H_0$  given an event  $E$ , which simultaneously yields the probability of the alternative hypothesis  $p(H_1|E) = 1 - p(H_0|E)$ . That is, Bayesian statistics evaluates the weight of the evidence from a study in support of one or the other hypothesis. An easy-to-use indicator for the strength of a statistical test is the Bayes factor (Rouder, Speckman, Sun, Morey, & Iverson, 2009).<sup>7</sup> Its power derives from the fact that it is not biased – as are  $p$ , effect sizes, and confidence intervals – in favor of the alternative hypothesis and therefore provides a measure for the quality of the evidence made for or against claims.<sup>8</sup> Tables that map calculated Bayes factors to qualitative expressions of the strength of evidence use a scale from “decisive,” “very strong,” “strong,” “substantial,” and “anecdotal” for both the null and alternative hypothesis (Table 1). Thus, a study that is statistically significant nevertheless may provide little more than anecdotal evidence for the hypothesis that there is an effect.

««««« Insert Table 1 about here »»»»»»

If we assumed for the moment that all of Bodzin’s tests are independent and calculated the Bayes factor based on the absence of prior knowledge (equal priors for null and alternative hypothesis), we would obtain the results in Table 1. These shows that 6 of the tests conducted provide only anecdotal evidence in favor of the alternative hypothesis and 4 tests provide anecdotal evidence in favor of the null hypothesis. As the implementation of one-tailed tests show, the author appeared to have had good reasons to anticipate positive treatment effects. Such prior beliefs may be used to adjust the statistics to account for prior knowledge. As soon as we assume that there is prior knowledge available in favor of larger effect sizes for the treatment, more of the tests become anecdotal evidence *against* the claims that the treatment applied by Bodzin *caused* the differences observed. Moreover, if we removed the overall test to avoid statistical dependence as well as the overall tests for each subscale, then there would be only four decisive tests left, three of which on the same (UHI) scale (Table 1)! Apart from one other test, the remaining evidence would be anecdotal only.

<sup>7</sup> Technically, the probability for the null hypothesis following the collection of data is given by  $p(H_0|E) = \left(1 + \frac{\pi_1}{\pi_0} \frac{1}{BF}\right)^{-1}$  where  $E$  denotes the data,  $BF$  is the Bayes factor, and  $\pi_0$  and  $\pi_1$  are the prior probabilities of  $H_0$  and  $H_1$ , respectively with  $\pi_1 = (1 - \pi_0)$  (Gonen, Johnson, Lu, & Westfall, 2005).

<sup>8</sup> One of the fundamental lesson beginners in statistics learn is that one “cannot prove or provide evidence *for* the null hypothesis.”

The upshot of this ever-so-brief analysis is that the evidence in favor of a treatment effect in Bodzin's study is rather weak and at best anecdotal – apart from being subject to the serious threats to the validity of the experiment deriving from the failure to exclude alternative explanations. Even if all this were not problematic, there would still be the question what the study says to science teachers and policy makers, an issue even for the best-constructed studies such as PISA (OECD, 2010). Thus, as Figure 1 shows, because the overlap between the two distributions is so large – i.e., within group variation ( $SD = 98$ ) large compared to between group variation ( $X_{\text{BOYS}} - X_{\text{GIRLS}} = 14$ ) – we do not know whether a particular girl or group of girls can be said to be doing better or worse than boys. Similarly if Figure 1 were to express the results of an experimental or quasi-experimental study, we would be unable to say whether a particular girl or group of girls had benefitted from the treatment because she/it achieved higher than some boys but lower than other boys (Ercikan & Roth, submitted). Frequency-based statistics therefore come with considerable limitations concerning the weight and interpretability of the evidence collected in a study. As a result, whether frequency-based statistics can provide useful recommendations to practitioners and policy-makers depends on the degree to which study findings apply to the relevant individual or subgroup of individuals.

### Conclusion

Science educators, as scholars in any other science, ought to strive to provide the strongest forms of evidence for the claims they make. For the evidence to be strong, the design of studies needs to rule out alternative explanations to the largest extent possible. This pertains to single (qualitative) case studies as to high-powered statistical work using the most advanced mathematical modeling techniques and experimental designs. Moreover, because there are many problems with traditional statistics, some substantial, science educators ought to choose the strongest possible statistical methods available to them. In the face of a public debate about evidence-based decision making in educational reform and in the face of efforts to make evidence-based reasoning itself a primary educational goal (e.g., Callan et al., 2009), science educators do not want to be the children left behind. We, science educators, owe it to ourselves to work together (authors, peer reviewers) to produce the strongest possible evidence in the construction of explanations.

### References

- Bakhtin, M. M. (1981). *The dialogic imagination*. Austin, TX: University of Texas Press.
- Behm, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100, 407–425.
- Bodzin, A. M. (2011). The implementation of a geospatial information technology (GIT)-supported land use change curriculum with urban middle school learners to promote spatial thinking. *Journal of Research in Science Teaching*, 48, 281–300.
- Bourdieu, P. (1992). The practice of reflexive sociology (The Paris workshop). In P. Bourdieu & L.J.D. Wacquant, *An invitation to reflexive sociology* (pp. 216–260). Chicago, IL: University of Chicago Press.

- Callan, E., Grotzer, T., Kagan, J., Nisbett, R. E., Perkins, D. N., & Shulman, L. S. (2009). *Education and a civil society: Teaching evidence-based decision making*. Cambridge, MA: American Academy of Arts and Sciences.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.
- Ercikan, K., & Roth, W.-M. (Eds.). (2009). *Generalization in educational research*. New York, NY: Routledge.
- Ercikan, K., & Roth, W.-M. (submitted). *Generalizing from educational research: The validity of evidence used to inform policy and practice*.
- Gonen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). *The Bayesian two-sample t-test*. Memorial Sloan-Kettering Cancer Center Department of Epidemiology and Biostatistics Working Paper Series. Working Paper 1. Accessed August 29, 2011 at <http://www.bepress.com/mskccbiostat/paper1>
- Gurrin, L. C., Kurinczuk, J. J., & Burton, P. R. (2000). Bayesian statistics in medical research: An intuitive alternative to conventional data analysis. *Journal of Evaluation in Clinical Practice*, 6, 193–204.
- Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach*. Chicago, IL: Open Court. (First published in 1989)
- Leontyev, A. N. (1981). *Problems of the development of mind*. Moscow, USSR: Progress Publishers.
- Organisation for Economic Co-operation and Development (OECD). (2010). *PISA 2009 results: What students know and can do. Student performance in reading, mathematics and science (vol. 1)*. Accessed June 29, 2011 at URL <http://dx.doi.org/10.1787/9789264091450-en>.
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16, 225–237.
- Stegmüller, W. (1974). *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie, Band I: Wissenschaftliche Erklärung und Begründung [Problems and results of the theory of science and analytic philosophy vol. 1: Scientific explanation and rationale]*. Berlin, Germany: Springer-Verlag.
- Vygotsky, L. S. (1997). *The historical meaning of the crisis in psychology: A methodological investigation*. In *The collected works of L.S. Vygotsky vol. 3: Problems of the theory and history of psychology (pp. 233–343)*. New York, NY: Plenum Press. (First published in 1927)
- Wagemakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, 100, 426–432.
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Sciences*, 6, 29–298.

### **Caption**

Figure 1. A plot of two population distributions representing science scores of U.S. boys and girls based on the means and standard deviation reported by the 2009 PISA study (OECD, 2010).