Journal of Scientific Exploration, Vol. 27, No. 1, pp. 5–11, 2013 0892-3310/13

EDITORIAL

Periodicals of various sorts have long recognized the need to address certain topics on a regular basis. That's why computer magazines routinely offer articles such as "Windows Tips and Tricks," and "How to Protect Your Data." Similarly, photography magazines return again and again to articles explaining how to get the most out of wide-angle lenses, how to shoot portraits in natural light, or how to photograph dramatic landscapes. It seems to me that *JSE* editorials might also need to recycle certain topics from time to time, in part because readership changes, and in part because researchers in areas of frontier science can have conveniently short memories (like everyone else), perhaps especially when it comes to matters that are intellectually or professionally challenging or uncomfortable.

The continuing debate over Daryl Bem's recent precognition experiments (see Bem 2011, and the Editorial in *JSE* 25:1) and the similar controversy still dogging work on LENR or "cold fusion" suggests that perhaps it's time to review certain salient facts about the nature of experimental replication in science. What follows is not new. Harry Collins has done outstanding work on this topic (Collins 1992), and I also addressed the issue at length (Braude 2002). For more recent commentary, see also Stefan Schmidt (2009). Apparently, however, what's both obvious and commonsensical is very easy to overlook.

Of course, it's clear enough why so much emphasis is placed on the replication of experiments, not just in parapsychological and LENR research but in other areas of science as well. Experimental replication would seem to be an obvious and straightforward means of legitimizing experimental results. The underlying idea is that if an experiment *E* gives a certain result while attempted replications do not, we have good reason to regard *E*'s result as spurious or inconclusive. And if continued attempts to replicate *E* fail to duplicate *E*'s result, we have (so the story goes) good reason for regarding the outcome of *E* to be due to a flaw in *E*'s experimental design, or to experimental negligence or incompetence, and perhaps even to chicanery. So the received view is that the only legitimate experimental results in science are those that can be repeated reliably, and in this way scientific repeatability has served as a kind of supplementary demarcation criterion (after falsifiability) between science and non-science (or pseudoscience).

I assume that nearly all *JSE* readers are familiar with this story. But I have to wonder how many of them realize that it rests on an unacceptably naïve conception of what experimental repeatability actually is, as well

as an even deeper conceptual confusion over the nature of similarity. The former is simply a special case of the latter.

The first point worth considering is that, despite considerable scientific posturing to the contrary, when it comes right down to it—especially in situations when the scientist's own work is on the line—experimental replicability *in fact* is rarely (if ever) considered to be an essential feature of genuine science. Rather, it's typically regarded as such primarily in politically charged debates over psi research, LENR, and some other areas of frontier science. In those debates, defenders of the replicability requirement (let's whimsically call them *replicants*) seem conveniently to forget, first of all, that criteria of (and reliance on) replicability vary considerably from one area of science to another. Not surprisingly, these differences are especially pronounced when we compare behavioral sciences to nonbehavioral sciences. But even in the physical sciences, the importance of (and reliance on) replicability varies greatly—say, from geology and astronomy (not to mention cosmology and meteorology) to physics and chemistry.

But a much more serious problem is that the very *concept* of experimental replication is exceedingly crude. To see why, let's begin by asking: In what respects can replication attempts differ from the original experiment? It's clear, first of all, that no replication attempt can ever be *exactly* the same as the original, if only because of changes in the time and place of the experiments. But of course those differences will be accompanied by differences in the general conditions of the experiment or in the experimental environment. And these may include differences in the actual participants. But even if the participants remain the same, we can expect changes in their attitude or mood, or even in the condition of the experimental apparatus required (especially sophisticated, sensitive, or delicate equipment), all of which might vary subtly or dramatically from one test to another.

Notice that—even in the "hard" sciences—these sorts of differences between experiments and their replication attempts are tolerated all the time (if they're noticed at all). In physics, an experiment conducted at laboratory L with a certain kind of particle accelerator might be replicated at laboratory L' with a different design of accelerator. In microbiology, experiments conducted with microorganism M in solution S might be replicated by studying M in a different solution S' (which may have been more convenient to use, but whose differences are considered insignificant). In fact, even a different microorganism M' might have been substituted and its difference discounted. And of course, despite the expectations of the replicating scientist, it's always possible that such differences between experiments lead to differences in experimental outcome. For example, in

physics, some of the differences between experiments and their attempted replications might account for the mixed results of efforts to test the EPR paradox and hidden-variable interpretation of quantum mechanics. In fact, these attempts didn't even study the same particles. One used proton pairs (McWeeny & Amovilli 1999), and the others, photons (e.g., Freedman & Clauser 1972, Aspect, Dalibard, & Roger 1982, Aspect, Grangier, & Roger 1982). Yet they're all considered versions of the same experiment,[1] originally proposed in a thought experiment by David Bohm, but which involved electron-pairs (Bohm 1952a, 1952b, Bohm & Aharonov 1957). At any rate, the standard procedure in cases such as this is to ignore the differences between these experiments so long as their results *more or less* agree, and thus to regard the follow-up experiments as replications of the earlier ones. But if the experiments produce sufficiently dissimilar results, the standard procedure is to regard the later experiments as failing to replicate the former.

There's a very important moral to this story. If we pay attention to the way the business of science is actually conducted, what we find is that criteria of experimental replicability are both very loose and never fully specified. In fact, scientists don't decide whether follow-up experiment E_2 counts as a replication of original experiment E_1 until the results of E_2 are in. It's certainly not decided solely on the basis of formal features of the two experiments—something potentially expressible in a "recipe" or unambiguous and complete list of all relevant procedures. On the contrary, when scientists agree that E_2's results match those of E_1, they will simply ignore the unavoidable and potentially relevant differences between E_1 and E_2, declare that E_1 has been replicated, and (in some cases) conclude that the results lend confirmatory weight to a shared, underlying, and trusted theory. But if E_2 fails to yield the hoped-for (and possibly only approximate) duplication of E_1's results, the standard reaction is to suppose that the inevitable differences between the two experiments in fact made a difference and that this failure does not automatically cast doubt on or discredit the original experiment's results or the shared underlying theory. As a rule, then, both avoidable and unavoidable differences between experiments and replication attempts are tolerated all the time, and ignored so long as the results pan out more or less as expected, but invoked when results go the other way.

Another way to put the point is this: Whether or not the differences between E_1 and E_2 count as relevant is *not* determined independently of the decision as to whether the latter replicates the former. Scientists tend to regard many such differences as important only if the outcomes of the experiments differ. But before knowing the results of E_2, it's pretty much an

open question whether the differences between E_1 and E_2 matter. Of course, scientists may claim in advance that the differences don't matter, but if the replication attempt fails to give more or less the same results as the original experiment, they may retract that judgment.

The situation changes somewhat when a series of replication attempts fails to consistently produce results similar to the original experiment. But even then (as we've seen recently with attempts to replicate Bem's experiments), the same general attitude about replicability prevails. When the later experiments fail to produce positive results like those obtained by Bem, the conversation focuses, for instance, on the differences in the protocols, or the different attitudes of the experimenters. And again, it's likely that these differences would also have been ignored had the later results all been positive. After all, some attempts to replicate Bem's experiments *have* been considered successful, and they're not strictly identical to the experiments Bem originally performed. Furthermore, there's nothing inherently suspicious or unsavory about this. That's simply the way science works, and given the inevitable differences between original experiments and replication attempts—magnified in the behavioral sciences by many additional kinds of potentially relevant variables—it's the only way it can work.

Interestingly, many consider replication attempts successful and convincing *only* when they're conducted by someone other than the original scientist. In part, I suppose, it's because they believe that any legitimate experiment can be described in a list of procedures which any competent scientist should be able to follow and produce the same results. For example (and somewhat notoriously), Karl Popper wrote, "any empirical scientific statement can be presented (by describing experimental arrangement, etc.) in such a way that *anyone who has learned the relevant techniques* can test it" (Popper 1959:99, emphasis added). This position is especially dubious when applied to parapsychology, alternative healing experiments, and the behavioral sciences generally, where experimenter expectancy effects and the variability of subject–experimenter interactions are particularly problematical. But it's also an obviously questionable position to take with respect to any area of frontier science, where the relevance of numerous and unavoidable differences between experiments hasn't yet been determined. In fact, I'd say that one of the most important lessons learned from the behavioral sciences, and reinforced by studies in many areas of frontier science, is that it's still an open question whether it's reasonable to expect success when replication attempts are conducted by someone other than the original experimenter. Moreover, it's unclear to what extent this might be an issue in mainstream science, where (as Rupert Sheldrake has noted (1998)), double-blind protocols are typically neither used nor even taught as sound

methodology, and where potential experimenter effects are not even on the radar.

As I mentioned above, some difficulties in determining when an experiment has been repeated are not peculiar to the scientific enterprise or to the process of experimentation. Rather, they're an instance of the more general problem of determining when *any* sort of event has been repeated. These problems, in other words, concern the general concept of *recurrence*, and even more fundamentally, the concept of *similarity*.

Suppose that *A* tells a certain joke and that his telling of the joke, *J*, is very funny. But suppose that *B*, who is not as comedically gifted as *A*, tries to tell *A*'s joke using different words, inflection, and timing, as a result of which his joke-attempt *J′* is not funny. How, then, do we answer the question: Is *J′* a recurrence of joke-attempt *J*? The important thing to observe here is that this question has no simple or straightforward answer. There are perfectly acceptable reasons for answering it either affirmatively or negatively. Some might say that although *B* told the same joke as *A*, he didn't do so with the same (or perhaps any) comedic skill. On the other hand, some might claim that, since *A* and *B* uttered different strings of words, and since *J′* was not funny, *A*'s joke had *not* been repeated by *B*.

The important point to grasp here is that neither response is intrinsically better than the other. Whether we take *B*'s performance to replicate *A*'s performance depends on what's appropriate for the context in which the question arises. Suppose people are taking turns telling jokes at a party and that each person is expected to tell a different joke. If *B* were to tell his joke, we might feel justified in complaining that he didn't tell a new joke and in fact that he merely told *A*'s joke rather poorly. On the other hand, suppose the party guests are playing a different game, in which each has to memorize and repeat verbatim what his immediate predecessor says. Suppose, then, that *A* tells his joke and that *B*, whom we may suppose is mnemonically challenged, tries unsuccessfully to repeat *A*'s performance. Even if the *content* of what *A* and *B* said was similar, so that we might consider *B* to have succeeded in producing a *version* of *A*'s joke, *B*'s performance (the string of words produced in the manner produced) would not count in this context as a replication of *A*'s performance. We can imagine even more stringent requirements of replicability. Suppose *B* is studying the comedic arts, and that his task is to repeat, *not* just the same words as those of his teacher *A*, but also *A*'s inflection and timing (and note, criteria of sameness for inflection and timing are not hard and fast; for example, we needn't suppose that *A* and *B* have voices of the same quality). In this context, what *B* does will not be a recurrence or replication of what *A* does, if *B* manages to get only the words exactly right.

The moral of all this is that whether or not *B*'s verbal performance constitutes a recurrence (or replication) of *A*'s joke-telling *J* is not simply a function of formal features of what *A* and *B* do and say. In one context *B*'s sequence of words might count as a recurrence of *J*, while in another it might not.

This is simply a real-life example of a point that applies even to the most elementary examples in mathematics, which likewise demonstrate that the relation "__is similar to__" is not simply a static, two-termed relation between things, but is inevitably tied to contextual and variable criteria of relevance that are not part of an absolute inventory of Nature's furniture. As I've noted many times, this can be easily illustrated by an example from geometry, although mathematicians typically use the term "congruence" rather than "similarity" (for a more elaborate discussion of this example, see Braude, 2007, Chapter 7). In any case, mathematicians know that in the absence of some specified or agreed-upon rule of projection, or function for mapping geometric figures onto other things, no figure is congruent with (similar to) anything else. They recognize that there are different standards of congruence, appropriate for different situations. Depending on which rule of projection we choose, we may consider a given triangle to be congruent only with triangles with the same horizontal orientation and the same angles, or we may consider it to be congruent with any triangle, or even with squares or lines. So in geometry, no property intrinsic to a given triangle determines which other geometrical figures that triangle is congruent with. And that's because no situation is *intrinsically basic*; standards of relevance emerge from living and ephemeral human situations, not from Nature herself. But then no standard of congruence or similarity is inherently privileged or more fundamental than others. And clearly, if this is true even for the comparison of simple geometrical figures, it's true *a fortiori* for the comparison of much more multi-faceted joke attempts and scientific experiments.[2]

~ ~ ~

A short but important note on a different matter. This issue contains a letter from Caroline Watt announcing the implementation of a webpage for registering parapsychological experiments. The value of this or any registry has recently been a hot topic for debate among parapsychologists, and, as far as I can tell, there's little consensus among researchers on the matter. Consequently, the *JSE* will remain neutral and allow researchers to decide for themselves whether to avail themselves of this opportunity to register their experiments. As a result, I feel it's important to note that the *JSE* will not require authors reporting parapsychological experiments to register their studies, and that registration will not be a factor in my editorial decisions.

Notes

[1] That's because (as James Spottiswoode was kind enough to remind me—personal communication) quantum mechanics "explicitly predicts that all these particles should show the same behavior. So failure to replicate across particles would have big consequences."

[2] I'm grateful to James Spottiswoode and Michael Ibison for some very helpful communications on the topic of this Editorial.

Stephen E. Braude

References

Aspect, A., Dalibard, J., & Roger, G. (1982). Experimental test of Bell's inequalities using time-varying analyzers. *Physical Review Letters, 49*(25), 1804–1807.

Aspect, A., Grangier, P., & Roger, G. (1982). Experimental realization of Einstein–Podolsky–Rosen–Bohm *Gedankenexperiment*: A new violation of Bell's inequalities. *Physical Review Letters, 49*(2), 91–94.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425.

Bohm, D. (1952a). A suggested interpretation of the quantum theory in terms of "hidden" variables. I. *Physical Review, 85*(2), 166–179.

Bohm, D. (1952b). A suggested interpretation of the quantum theory in terms of "hidden" variables. II. *Physical Review, 85*(2), 180–193.

Bohm, D., & Aharonov, Y. (1957). Discussion of experimental proof for the paradox of Einstein, Rosen, and Podolsky. *Physical Review, 108,* 1070–1076.

Braude, S. E. (2002). *ESP and Psychokinesis: A Philosophical Examination* (revised edition). Parkland, FL: Brown Walker Press.

Braude, S. E. (2007). *The Gold Leaf Lady and Other Parapsychological Investigations*. Chicago: University of Chicago Press.

Collins, H. M. (1992). *Changing Order: Replication and Induction in Scientific Practice*. Chicago: University of Chicago Press.

Freedman, S. J., & Clauser, J. F. (1972). Experimental test of local hidden-variable theories. *Physical Review Letters, 28*(14), 938–941.

McWeeny, R., & Amovilli, C. (1999). Locality and nonlocality in quantum mechanics: A two-proton EPR experiment. *International Journal of Quantum Chemistry, 74*(5), 573–584.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. New York: Harper.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90–100.

Sheldrake, R. (1998). Experimenter effects in scientific research: How widely are they neglected? *Journal of Scientific Exploration, 12*, 73–78.