Sci & Educ (2008) 17:547–558
DOI 10.1007/s11191-007-9111-9

Studies in Historical Replication in Psychology VII: The Relative Utility of "Ancestor Analysis" from Scientific and Educational Vantages

Michael Andrew Ranney

Published online: 15 September 2007

Abstract This article discusses, from various vantages, Ryan Tweney's (this issue) pedagogical technique of employing historical replications of psychological experiments with graduate students in psychology. A *prima facie* perspective suggests great promise for this sort of academic "ancestor analysis," particularly given the enthusiasm and skill represented in the activities that culminated in the replicators' articles. It is suggested that such activities might be enhanced by requiring a contextualization that makes contact with more modern psychological research—particularly regarding expositions of the replications. From a scientific/cognitive methods perspective, the original experimenters' inexplicit, ambiguous, descriptions provide both challenges and opportunities for students seeking to improve their understandings of their field. Three practical questions are posed herein regarding the general utility of this—or any—proposed instructional intervention. Ultimately, determining and integrating the diverse objectives that essential stakeholders have in graduate psychological training represent critical prerequisites in comprehensively assessing the relative advantages of such historical replications with respect to alternative experiences.

1 Introduction

This manuscript represents a discussion of the historical replication venture that Ryan Tweney and his graduate students (this issue) have engaged in. Since the authors of the preceding articles take history seriously, as do I, here is the micro-history of my involvement: I served as a discussant at their symposium at the 2005 CHEIRON conference (i.e., the 37th annual meeting of the International Society for the History of the Behavioral and Social Sciences), and was asked to, in part, provide a pedagogical perspective on the utility of historical replications of psychological research. I was initially

M. A. Ranney (✉)
Graduate School of Education, University of California, Berkeley, 4655 Tolman Hall, Berkeley, CA 94720-1670, USA
e-mail: ranney@cogsci.berkeley.edu

skeptical about the venture, and feared that my comments might be too "tough." However, after a meeting with Tweney and a good number of the students involved in the replications prior to the conference (in Bowling Green) many of my concerns were allayed; even more concerns were assuaged during the conference and by the end of the symposium. Still, some challenging questions lingered, which I will explicate in what follows. These questions are wide-ranging because they include ones that are virtually as abstract as science in general, as concrete as the particular character of psychological research, and as specific as involving ways to assess systematically aspects of the relative pedagogical utility of Tweney's historical replication technique.

In many respects, Tweney's introduction to the preceding articles points the way by highlighting how the nature of a field's methodology represents both the crux of its identity and a critical component of a field's status. In this discussion, my analysis of relevant methods takes on three heterogeneous manifestations: First, I will address historical replication as a pedagogical method from a somewhat informal, *prima facie*, perspective. Second, some features and caveats about the methods of the replications themselves will be described, and I will offer some observations regarding what we can learn about how scientific fields (whether seen as "hard" or "soft") are related. Third, methods for more fully determining the utility of historical replications as an instructional tool will be considered; this final analysis will involve a trio of its own—a triad of epistemic questions.

2 Why Historical Replications "Feel Right"

As Tweney (this issue) points out, there is considerable reason, even a priori, to believe that historical replications would be useful to psychology's graduate students. For instance, the "hands on" element is tantalizing. Indeed, a hallmark of one of the programs that I am most connected to at Berkeley—the Education in Mathematics, Science, and Technology (EMST) program—requires even first-year graduate students to carry out research, with project reports that are expected on the first day of both their second and third years. Most of these projects have a hands-on character, even though their instruments do not quite possess the "brass and glass" (Tweney 2003)—or perhaps, in the context of the present replications, the "skull and sleep deprivation"—character of the projects that Tweney and his students engaged in during their homage to 1900-ish psychology. However, my reading of the research on educational "manipulatives" suggests that advocates of hands-on instruction often overstate what is warranted by existing data. Further, as will be elaborated upon more below, even if a technique (such as a hands-on one) yields a pedagogical benefit, one ought still consider the benefit relative to the effort required to obtain it—and relative to the benefit and efficiencies of other techniques that might have been, but were not, employed. In other words, although wrestling with the down-and-dirty of doing real experimentation doubtlessly offers rewards, sometimes the "cleanliness" of abstractions, such as a student merely exploring simulations of psychological experiments, may be superior (e.g., in time spent) in facilitating learning and long-lasting comprehension.

From the *prima facie* vantage, some of the observations that I made before, during, and after the CHEIRON symposium were more related to affect than to learning. For instance, of the ones I met in person, the "replicators" (as I will call the graduate students involved) largely seemed to think of the venture as a bit of an *ad*venture—that is, an enjoyable "kick." The historical replications offered more than a bit of fun and some fine cocktail-party fodder, especially for the sleep-deprived replicators and the recently minted "phrenologists." Such benefits are hardly trivial; a critical component of a psychologist's identity is a

sense of kinship to past researchers and a positive feeling of connectedness to prior paradigms—so the graduate students' faculty advisors would be gratified to learn that Tweney helped the replicators bond with, and thus further appreciate, their field. After all, a rather common finding is that people tend to like the topics that they know the most about. (Indeed, although one might question the direction of causality, I certainly enjoy the history of psychology much more now than when I took my first course in it!) Further, most of us are aware of students and colleagues who delight in recounting their "academic ancestry," and psychology is no exception—whether they see it descending from Freud, Pavlov, Piaget, Vygotsky, Watson, or Wundt. Senses of lineage and intellectual heritage, especially when synergistic with the historical spirit and the advancement of one's technical abilities in empirical science, are not to be dismissed; they seem to represent notable products of Tweney's instructional method.

This affective component being noted, though, I worry that not all of the replicators enhanced their mastery, scholarship, and sense of ownership regarding their "day job" research material; this perhaps varied with how closely their prior interests matched their particular replication. For instance, I note that Sirrine and McCarthy (this issue) offer little-to-no modern scientific contextualization for their analysis of Gertrude Stein's automatic writing research. Yet there exists decades of more recent (particularly cognitive) work from the areas of attention, psycholinguistics, and neuroscience, etc., that could bear upon the degree to which people could engage in (or Stein and her colleagues/participants could have engaged in) automatic writing. Perhaps, however, this relative focus on the historical instead of the more currently scientific was intentional; I note that Fuchs and Burgdorf's (this issue) sleep deprivation replication includes at least one author who indeed studies sleep processes—yet modern sleep citations are few in their paper as well. In the end, one can imagine that the utility of a historical replication might have an inverted-U sort of relationship to the match between a graduate student's central (day-job) research thrust: Too little fit might fail to engage and exercise a student adequately, whereas an extremely close fit might inhibit desirable knowledge elaboration and methodological "cross training."

So, as a friendly amendment to the historical replication approach that Tweney and his students (this issue) have pioneered, I would recommend explicitly adding (if it is not already required) a "modern literature perspective" section to their course reports. For instance, McConnell and Konrad's article (this issue) might have addressed, space permitting, modern notions of neural localization/lateralization or delocalization—notions such as the "new phrenology" of fMRI or single-cell recording research (e.g., Christof Koch's "Jennifer Anniston cells;" Faw 2005), the relationships between brain structures and spatial or other cognitive abilities (e.g., Burgess et al. 2002), or even suggestions of the relative modularity of mind regarding language faculties (e.g., Fodor 1983). Similarly, Fuchs and Burgdorf (this issue) might have included connections (which I know they are aware of) between their research and modern studies of vigilance—such as the sorts of "microsleep" phenomena observed among long-haul jet pilots and others (Dinges 1995; Mumaw, personal communication, March 26, 2006; Rosekind 2005). Further, Ayala et al. (this issue) could relate Wundt's metronome-triggered phenomena to more modern conceptualizations of sensory, short-term, and long-term memory—or even the chunking or grouping processes studied by Miller (1956), Herbert Simon and colleagues (e.g., Richman et al. 1995), or researchers from the Gestalt School (cf. Palmer 2003; also see Bregman 1990, on grouping specifically in the auditory domain). Following this thread to Athy et al.'s article (this issue), the Brunswikian contrast of perception with explicit reasoning could find many more modern parallels in problem solving research—for instance, research on the nine-dot problem (e.g., Chronicle et al. 2001, regarding realistic

magnitudes from the replication; also see Paige and Simon's 1966 work on algebra, in which participants generated proposed solutions that were physically impossible).

A potential pitfall of this suggested amendment, unfortunately, is that a replicator might be tempted simply to conclude that earlier research was poor, naïve, or unsophisticated. If implemented, then care would be required to balance such a potential for bias with a more charitable view of psychologists past—including the prescription that modern researchers should beware of hubris and only partially warranted pride. Part of being a graduate student is realizing that what you thought were your novel and brilliant ideas/distinctions may be, as the psychologist (and economics Nobel laureate) Herbert Simon would refer to them, "old wine in new bottles"—that is, constructs already entertained by intellectual progenitors such as Skinner, Lewin, (Max or Michael) Wertheimer, Kant, Hume, and/or even Plato.

3 "Though this be Madness, Yet there is Method in 't"

This well-known quote from Shakespeare's *Hamlet* is meant to highlight the zany, the ill-specified, and certainly the clever aspects of the diverse attempts at following the then-prevailing scientific methods that are represented in the original experiments taken on by the replicators. Part of what makes replications such as these of interest is they bring to the fore (a) how difficult it is to "do good science," (b) how difficult it is to measure cognitive constructs, (c) the virtually infinite regression of methodological explication that is possible, (d) how often methodology is implicitly undersold, and (e) that even failed or sub-veridical replications may provide useful understanding.

Again, Tweney's introduction (this issue) provides a helpful "hook," as he describes the method-driven fragmentation of psychology (which actually seems quite modest, relative to that in the field of education; cf. Labaree 2004). If anything, I believe that the historical replications suggest that the brass-and-glass days of psychology represented an immature discipline cloaked in the trappings of "hard science" technology. One could have had finely machined planchettes and skull-calipers, but such devices, without the advent of the theory-driven methods that attend statistical analysis, scarcely settled the issues of automatic writing or phrenological diagnosticity. Even today, the array of brain-imaging devices available is of little utility without their associated quantitative methods.

In addressing concerns relating to "methodological idolatry," Tweney noted how extremely central statistical coursework has become in psychology's graduate training. If, as Tweney invoked, "statistics makes us one"—namely a scientific psychologist—then I'm proud to be in that number because I am a scientist and there are much worse scientific paraphernalia with which to be associated. As is readily apparent, what has whittled away at the prominence of (e.g., machinists') "shops" in psychology departments—which made the brass, glass, wood, and plastic, etc., apparatuses of the past—is the development of our general-purpose computers. Because computers now make delivering most stimuli so easy, and because statistics also essentially reside in computers (e.g., the raw, transformed, and inferential data from imaging studies), the ascendant power of statistics (for getting at measuring constructs, etc.) is all the more palpable in contrast to the decline of the physical apparatus.

Statistical utility, as highlighted by the replications, permeates not only the graduate curriculum. It becomes central for teaching college students, as well. In my undergraduate Introduction to Cognitive Science course, for instance, I implicitly make statistical utility salient by asserting that the most fundamental problem facing cognition is that of figuring out how similar two entities are, in psychological terms. Of course, this is a classic

measurement problem—and thus a methodological problem—and the meaning(s) of similarity largely still vexes us after more than a century of study. In general, cognitive science is reasonably good at describing the similarity of the same pure auditory tone at two different volumes, but measuring psychological accuracy (or error) among heterogeneous real-world quantity estimation tasks remains challenging (Ranney et al. 2001; cf. Brown 2002, and Munnich et al. 2007). Further, we are clearly not yet close to describing precisely how similar a banana is to an apple (not to mention vice versa!), and we are quite far from pinning down how similar the *concept* of a banana is to the *concept* of discrimination (e.g., discriminating among fruit). If the present replications show nothing else, they show how hard it is to measure psychological constructs (e.g., the components of consciousness, in Wundtian fashion; Sahakian 1968; Wertheimer 1987), and this is a valuable lesson for psychology graduate students to learn first-hand as soon as possible. Indeed, this is even part of EMST's "fail early, fail often" (partially tongue-in-cheek) philosophy of graduate training.

Pulling back from the data analytic perspective, we see another common theme: the replicators' descriptions of their activities often involved frustration (albeit not without bemusement) regarding how inexplicitly and ambiguously the original scientists documented their methods. As it does today, of course, such explication varied with the particular genre of writing that a replicator was examining (e.g., an extended abstract vs. an experimental journal article)—and extremely explicit prose risks becoming mind-numbing for a reader. But it does seem that, on the whole, psychology has advanced considerably in terms of explaining the precise nature of its methods and analyses. In contrast, the "state of the art" of the bulk of the original experiments that the replicators took on would hardly be accepted (as one would hope) into august journals such as *Psychological Review* today. Still, one wonders how charitable readers will be with the current body of research a century from now. It could be that there are implicit "you know what I mean" phrases throughout our Method sections that will mean as much to future scientists as "tachistoscopes" (McClelland and Rumelhart 1981)—let alone "gravity chronographs" (e.g., Cattell 1885)—would mean to current sets of first-year graduate students.

It is not uncommon to find psychological or other journals that print Method sections in smaller lettering than the surrounding text. Why is that? Regardless of the original reasons, the overall message about the need to read the section is hardly lost on the articles' audience—including the journals' current and aspiring contributors. Like most such decisions, the "font diminution" choice is likely the result of multiple causes. But among the reasons is undoubtedly that we convince ourselves that the methods are probably quite similar to, and thus semi-redundant with, those described in other papers by the same or associated authors.

As it turns out, none of the replicators could certify that fully veridical replications of the original experiments were carried out. In some instances, the manuals or method sections (as such) were inexplicit, leaving the replicators to infer or invent the interstitial techniques, such as techniques on how to train (e.g., Wundtian) participants. In others, such as in the sleep deprivation replication, only a partial replication was performed (as Fuchs and Burgdorf (this issue) report that they employed somewhat fewer hours and assessments). Even so, a historical appreciation may arise from such sub-veridical replications. Regarding the less flattering side of the psychological heritage, the "Lake Wobegon" effect[1] that replicators McConnell and Konrad (this issue), in essence, report about our

[1] The effect refers to purported cases in which members or subsets of a population are implausibly "above average" (Cannell 1988; Koretz 1988; referring to Keillor's 1985, etc., stories).

phrenological predecessors' results makes contact with modern elements of social desirability. Along with the phenomena of horoscopes, tarot-reading and palm-reading, showing that "all phrenology clients are above average" suggests that not all of the researchers who came before us were above some self-serving self-delusions or were even above "little frauds." On the more uplifting end, Patrick and Gilbert's (1896) work still looks remarkably good, as Fuchs and Burgdorf (this issue) point out. Such research, as with much seminal work by innovators such as Ebbinghaus (1885/1913) or Bryan and Harter (1887), *do* permit us to consider a number of these "long dead forebear folk" as giants.

4 Pedagogical Science and Criteria for Evaluating Historical Replications

From the standpoint of pedagogy, I generally assess extant (or even proposed) instructional techniques with litmus-test questions that are reminiscent of how one might use the magazine *Consumer Reports* to decide about a purchase: (Q1) How does a proposed focal pedagogical intervention compare with other such interventions (existing or plausible) that might be oriented toward achieving the same goals? (Q2) How do these possible pedagogical choices compare with having no intervention at all (that is, compared to the "spontaneous remission" of ignorance), relative to the cost of the intervention—in both time and other resources? (Q3) How long lasting are these interventions' effects? As one's lifetime is not (yet) infinite, educational researchers rarely address all of these three questions when they propose some instruction. Each of these questions is difficult to ask one's self as a researcher, as each strikes at the heart of our confirmation biases—and perhaps thus at the "imposter syndrome" (e.g., Kolligian and Sternberg 1991) that may keep us from attempting disconfirmations of our most treasured hypotheses (see Tweney et al. 1981, on disconfirmation attempts). In this section, I will address Q1, Q2, and Q3 rather abstractly, and follow that by more concretely considering historical replication—as a pedagogical technique—with respect to each of the questions.

When dealing with prominent alternative instructional techniques, finding a good, honest, answer to the differential utility question, Q1, requires no small amount of resources; therefore, such situations are often part of politically charged "big science." However, when a technique breaks new ground, there may be few, if any, contrastive alternatives that address the same educational goals in reasonably related ways. (For Berkeley's Reasoning Group, this was essentially the case for the ConvinceMe system with its ECHO reasoning engine, e.g., Ranney and Schank 1998—and is the case with curricular elements of Numerically Driven Inferencing: e.g., Munnich et al. 2004; Ranney et al. 2007.) Therefore, "finding comparables" is sometimes problematic, and this seems likely for Tweney's (this issue) technique.

Proponents of particular interventions commonly answer the first part of Q2—that is, the question of comparing something to nothing; but in the absence of Q2's second part, involving various costs, such efforts smack of straw-person hypothesizing or simply further replications of the time-on-task effect. (In the vernacular of one undergraduate, time-on-task replications are suitable for publication in the *Journal of "Duh!"*) A major problem with educational research is that new techniques often come in with considerable enthusiasm, a wheelbarrow of dollars, and many hours of researchers' (and/or teacher-release) time. But when the party is over, researchers usually focus on the finding of significant effects and worry less about whether the size of the effects warranted the funding and (researcher, instructor, or student) person-hours involved. Practitioners (e.g., classroom teachers) who are usually without extra resources naturally view some reports of learning

effects as what I call "brochure science"—and therefore reasonably discount them. Therefore, particular attention to the cost portion of Q2 suggests that minimalist interventions that provide good "bang for the buck"—or "bang per instructional minute"—learning experiences are noteworthy (e.g., from our laboratory: Garcia de Osuna et al. 2004; Munnich et al. 2005; Rinne et al. 2006). At the extreme, even a very short pause can represent an important "intervention" (e.g., Fox 1991; Tobin 1987).

The longevity, or "half-life," question (Q3) is similarly commonly forgotten in the rush to report promising results. This is unfortunate, given that delayed post-tests are often easy to implement (e.g., Ranney and Schank 1995; Munich et al. 2005) and can ease stakeholders' fears about the "out of sight, out of mind" alternative hypothesis that an effect may be rather transient.

5 How Might Historical Replications Fare Regarding the Three Questions?

Given some of the potential (or even likely) benefits of Tweney's technique (e.g., improved comprehension of, appreciation of, and attachment to psychology), consider the three criteria in light of the replication pedagogy. Questions Q1 and Q2, regarding the relative value of intervening, effectively makes contact with the notion of a marketplace of educational products, of which historical replication is a possible "consumable" in a "niche" market. The particular niche for Tweney's technique is that of graduate students in a history of psychology course, so were one to assess scientifically the strategy's utility, one would likely want to do something like controlled experiments in which sufficient numbers of such students would also be randomly assigned to courses using the traditional (e.g., lecture) format to address Q1, or assigned to no course at all to address Q2. (Naturally, it is difficult and/or unethical to provide a "placebo" curricular format.) To address Q3, regarding effect longevity, one probably would want to follow up a pre-test/post-test design for each of the three types of groups (experimental, traditional, and no-treatment) with delayed, longitudinal, post-tests at reasonable intervals—say after 1 year, 5 years, and perhaps even 30 years, assuming nonlinear effects.

Of course, carrying out an experiment to address Q3 along with Q1 and/or Q2 would not be easy. Q2 is problematic, given that Tweney's course is a doctoral requirement; it would be questionable to waive the course for a no-treatment group without those students—and their peers—knowing that is was waived.

Indeed, even if one eliminated the no-treatment group to address Q1, challenges would remain, regarding the two remaining—experimental and traditional—groups: (a) students within the institution would probably contrast their experiences across the two courses, (b) it would be difficult to keep the instructors naïve regarding the hypotheses, and (c) it would likewise be difficult to equate for time-on-task and instructor experience, etc. Multi-site and multi-instructor designs might be employed to assess the historical replication technique, but that would represent a rather massive, complicated, undertaking for a relatively niche market; these concerns relate to current, and often politicized, discussions about "evidence-based education" and the relative importance of randomized clinical educational trials.

An anonymous reviewer of this article suggested an alternative to gain purchase regarding Q1: use a within-course manipulation such that mutually exclusive complements of students tackle a topic with or without replications. A related empirical design, which I have recently deployed in trying to improve the numeracy displayed by graduate students in journalism, involves delaying the curricular module within a semester for the "control"

group. The design involves pre-, mid-, and post-tests, such that the mid-test represents a post-test for the early group and a redundant pre-test for the late group (and thus a control for test/retest effects; Ranney et al. 2007). None of these designs are perfect for the present purposes, though—partly because Q3's assessment horizon is arguably so long in the case of replicating historical psychological experiments. In the case of journalists' numeracy, end-of-semester improvements in several straightforward measures—say, in calculation, in numerical attitudes, and in estimating socially relevant statistics—are noteworthy and indicative (However, some other, less immediate, measures of cognitive change are important and desirable in the realm of journalistic numeracy, too—e.g., developing numerical skepticism without falling into cynicism). As noted earlier, though, psychological-historical experimental replications entail goals that span decades; Tweney is presumably more concerned with long-term, and often more subtle, effects than with whether a student understands (or enjoys) some minutiae of phrenology after a multi-week experience with it.

Continuing this focus on Q3, difficulties may arise in (say, an independent panel) agreeing on which assessment measures would be employed: How much would they focus on (among other dimensions) affect, content knowledge, methodological knowledge, and success/engagement in one's field—and how much would "consumers" such as faculty advisors weight such measures in an overall index of course success (in the spirit of Brunswik and his student, Kenneth Hammond, both of whom the replicators Athy, Friedrich, and Delaney, this issue, cite)? (In keeping with the "ancestor analysis" theme of this article, I feel compelled to note that Hammond was an important mentor to me when I was an undergraduate.) With all of these complications, one can hardly find fault with Tweney for not (at least not yet) running a grand assessment of the historical replication technique.

From a different slant, replication is not as uncommon or niche-like *generally* as it is for the history of psychology. For instance, in virtually all of the scientific labs I experienced as a student—whether in chemistry, physics, biology, astronomy, psychology, etc.—I often replicated variations on "hall of fame" experiments from the past (Indeed, I employ such labs in the large cognitive science course I teach). However, my fellow students and I used more modern "brass and glass," instead of equipment from the 1500's or even early 1900's—and a central focus of many of the "hard science" labs I performed was to "get the right outcome," which is hardly the pedagogical intent of Tweney's replication technique.[2] Still, by analogy, given that quasi-historical labs are replete in coursework from kinematics to cognition, one would expect considerable utility from (less quasi-) historical replications of pre-1950 psychology. But this is still an empirical question; it is not inconceivable that the time a graduate student spends on a replication—or even spends on taking a psychological history course at all (as there remains some controversy about it)—might be time better spent running that novel, killer, experiment that will yield a superior post-PhD position. This might be a particularly realistic concern regarding an activity that most regard as pseudoscience, such as phrenology. Then again, future career/happiness timelines are not necessarily monotonically increasing, and there are many scenarios in which one will ultimately be better off with a richer understanding of history than with a ground-breaking experiment on one's *curriculum vitae* upon graduation.

[2] In my introductory cognitive science course, my teaching assistants ("graduate student instructors") and I take a middle path with our behavioral (as opposed to computational or other) labs; we try to focus on the processes, theories, and epistemology of empirical cognition, rather than on replicating (sometimes delicate) phenomena. Naturally, were a semester to go by without a single class replication bearing a resemblance to the results from its original publication, students would begin to wonder if they had stumbled into a course akin to phrenology!

As I tell my students, history rarely offers a proper control group, and the case of Tweney's (this issue) technique is no exception. Indeed, every time I teach a course I tinker with it and treat it, out of necessity, as an exercise in formative, rather than summative, assessment; would that faculty had the luxury of evaluating each of the myriad of instructional choices we make! In practice, we must often "fly by the seats of our pants," because if education is to cognition as engineering is to physics, then one thing educators lack (to continue the aeronautical analogy) is a reliable wind tunnel. At the practical level, our test beds are considerably more complex than those that analyze jet-wash and wind shear phenomena (which are perhaps comparable to social interactions and catastrophic learning disruptions in an educational milieu). Given the complexity, our models (learning theories) are also less precise, I suggest, than those in aviation (e.g., computational wind tunnel simulations). Furthermore, we cannot "re-set" a student's mind for another curriculum in the way we can often easily re-test the same fuselage at a new wind speed; each student's experience is a precious entity that we cannot treat merely as "sampling with replacement"—as one might in other fields (For instance, using a factorial experimental design for each curricular decision would quickly exhaust all available participants if one wishes to analyze the myriad of features a curriculum subsumes, such as the features one would find in an intelligent tutoring system; Merrill et al. 1992, etc.). Such facets make educational applications of cognitive science markedly more problematic than applications of most other sciences (even dwarfing other contemporary complex scientific applications—such as those regarding the roles of fat in a person's diet).

In a similar vein, we cannot *precisely* determine how much Tweney's 25 students learned about, or grew to respect, psychology as a result of their historical replications—relative to what they would have gained otherwise; modeling the integration of individuals' brains, histories, and new experiences is obviously a very long-term project that is not for those easily discouraged. I think it highly laudable, though, and a testament to Tweney's technique, that roughly half of the 25 initial students in his class were so enthused by their projects to have brought them to their current states of fruition, even through a voluntary second seminar—and considerable work beyond that. In the end, this zeal may well represent one of the most striking of assessments.

6 Conclusions

At the CHEIRON symposium, I confessed that, although I had (and still have) great appreciation for history in general, and considerable affection for the history of psychology, I wasn't entirely clear about what it meant for the replicators to, as was intended, learn what a "proper historical question" is (cf. Johnson-Laird 1988, p. 13, on causation and controlled experiments versus "the historian's dilemma").[3] Furthering my suspicions about the centrality of history in understanding science, some research from my own laboratory has found that laypersons'—and even professionals'—definitions and uses of fundamental historical/scientific[4] terms such as "evidence" and "hypothesis" are often

[3] Even so, I have always valued the utility of historical replication for helping to explain what happened during an episode—for instance, in 19th-century battlefield analyses.

[4] Historical and scientific research both employ evidence and hypotheses, with evolutionary studies representing an interesting blend of history and science. Most sciences, however, include elements of prediction and control, whereas historians generally shy away from these realms, keeping mainly to more basic analytic elements such as description and explanation (which sciences also include; Ranney 1998—as part of Kaufman et al. 1998).

surprisingly labile and context dependent (e.g., Ranney et al. 1996). (We subsequently found that part of what distinguishes evidence from hypothesis in one's mind is that evidence is, among other things, both less causal and—perhaps not surprisingly—more about the past than about the future; Diehl et al. 1999.) The lack of more than a feature-based, family-resemblance kind of distinction between evidence and hypothesis is problematically compounded by the often-spotty historical record typical in the research that might be replicated (see above). This further complicates assessments of the cognitive utility of historical replications—in particular, regarding students' understandings of what was done and why it was done (e.g., whether a particular technique was used truly for theoretical reasons or was, rather, rationalized for convenient or practical reasons).

Two of my more popular graduate offerings are entitled (a) "Getting Your Doctorate and Getting a Good Job," and (b) "Problem Solving and Understanding." Considering the "lessons" I have learned from teaching this pair of courses, and considering my *Consumer Reports* orientation from above, if my daughter were facing a historical replication as a student in Tweney's course, I would likely advise her with the following, backward-working problem-solving strategy: "First, develop a sense of the kind of job and the kind of lifestyle you would most prefer after grad school. After that, imagine what experimental replication would most likely help you achieve those collective goals. (Then come and talk to me some more, of course.) Finally, propose that replication to Professor Tweney, and do as much or as little work on the project as is appropriate to get the most out of the experience." Were this advice implemented well by a recipient, it would implicitly invoke the recipient's model of the replication with respect to the three questions I posited: "Might *certain* empirical replications be better for me than others (Q1)? Will implementing this replication with vigor be meaningfully better for me than slacking off (which would approximate doing nothing; Q2)? Will this experience stay with me for a reasonable period (Q3)?"

In this article, I have tried to tease out multiple perspectives on assessing the utility of historical replications in psychology, particularly as implemented by Tweney (this issue) and his cadre of intrepid graduate students. The students' papers clearly show that a great deal of learning took place both while performing and writing about the replications. But as I've laid out above, a proper, overall, evaluation of the venture would not only (likely) be a massive empirical undertaking, but it would require conversations with diverse stakeholders as to what the ultimate aims of the replications are. For those who might find this "evaluative avoidance by caveat"—well, that is also a time-honored part of the history of psychology. Besides, I just offered above how I would advise my own child; that's a lot more definitive than one usually gets from those who study mental processes!

Acknowledgments I thank Ryan Tweney, Luke Rinne, Ed Munnich, Andrew Galpern, Patricia Schank, Janek Nelson, Jed Stamas, Myles Crain, Lauren Barth-Cohen, Kelvin Chan, Nelson Bradley, Tammie Chen, Sarah Cremer, Barbara Ditman, Luke Miratrix, Michelle Million, Rachel Ranney and anonymous reviewers for their comments on drafts of this manuscript.

References

Bregman AS (1990) Auditory scene analysis: The perceptual organization of sound. MIT Press, Cambridge, MA

Brown NR (2002) Real world estimation: estimation modes and seeding effects. In: Ross B (ed) Psychology of learning and motivation: advances in research and theory, vol 41. Academic Press, New York, pp 321–359

Bryan WL, Harter N (1887) Studies in the physiology and psychology of the telegraphic language. Psychol Rev 4:27–53

Burgess N, Maguire EA, O'Keefe J (2002) The human hippocampus and spatial and episodic memory. Neuron 35:625–641
Cannell JJ (1988) Nationally normed elementary achievement testing in america's public schools: how all 50 states are above the national average. Educ Measure Issues Practice 7(2):5–9
Cattell JM (1885) The inertia of the eye and the brain. Brain 8:29–312
Chronicle EP, MacGregor JN, Ormerod TC (2001) When insight just won't come: the failure of visual cues in the nine-dot problem. Q J Exp Psychol 54:903–919
Diehl C, Ranney M, Lan G, Castro S (1999) Hypotheses and evidence about evidence and hypotheses. In: Hahn M, Stoness SC (eds) Proceedings of the twenty-first annual conference of the cognitive science society. Erlbaum, Mahwah, NJ, p 791
Dinges D (1995) An overview of sleepiness and accidents. J Sleep Res 4:4–11
Ebbinghaus H (1913) Memory: a contribution to experimental psychology. In: Ruger HA, Bussenius CE (Trans) Teachers College, Columbia University, New York (Original work published in 1885)
Faw B (2005) What we know and what we don't about consciousness science: a review of ASSC-9 at Cal Tech, June 24–27, 2005. J Conscious Stud 12(7):74–86
Fodor JA (1983) Modularity of mind: an essay on faculty psychology. MIT Press, Cambridge, MA
Fox BA (1991) Cognitive and interactional aspects of correlation in tutoring. In: Goodyear P (ed) Teaching knowledge and intelligent tutoring. Ablex, Norwood, NJ, pp 149–172
Garcia de Osuna J, Ranney M, Nelson J (2004) Qualitative and quantitative effects of surprise: (Mis)estimates, rationales, and feedback-induced preference changes while considering abortion. In: Forbus K, Gentner D, Regier T (eds) Proceedings of the twenty-sixth annual conference of the cognitive science society. Erlbaum, Mahwah, NJ, pp 422–427
Johnson-Laird PN (1988) The computer and the mind: an introduction to cognitive science. Harvard University Press, Cambridge, MA
Kaufman D, Ranney M, Ohlsson S, Reiser B, Shapiro L (1998) Multidisciplinary perspectives on evolutionary reasoning [symposium summary]. In: Gernsbacher MA, Derry SJ (eds) Proceedings of the twentieth annual conference of the cognitive science society. Erlbaum, Mahwah, NJ, p 10
Keillor G (1985) Lake wobegon days. Viking, New York
Kolligian J Jr, Sternberg RJ (1991) Perceived fraudulence in young adults: Is there an imposter syndrome? J Pers Assess 56:308–326
Koretz D (1988) Arriving at Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? Am Educ 12(2):8–15, 46–52
Labaree DF (2004) The trouble with ed schools. Yale University Press, New Haven
McClelland JL, Rumelhart DE (1981) An interactive activation model of context effects in letter perception: Part 1. An account of basic finding. Psychol Rev 88:375–407
Merrill DC, Reiser BJ, Ranney M, Trafton JG (1992) Effective tutoring techniques: a comparison of human tutors and intelligent tutoring systems. J Learn Sci 2:277–305
Miller GA (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 63:81–97
Munnich EL, Ranney MA, Appel DM (2004) Numerically-driven inferencing in instruction: the relatively broad transfer of estimation skills. In: Forbus K, Gentner D, Regier T (eds) Proceedings of the twenty-sixth annual conference of the cognitive science society. Erlbaum, Mahwah, NJ, pp 987–992
Munnich E, Ranney M, Bachman MLN (2005) The longevities of policy-shifts and memories sue to single feedback numbers. In: Bara BG, Barsalou L, Bucciarelli M (eds) Proceedings of the twenty-seventh annual conference of the cognitive science society. Erlbaum, Mahwah, NJ, pp 1553–1558
Munnich EL, Ranney MA, Song M (2007) Surprise, surprise: the role of surprising numerical feedback in belief change. In: McNamara DS, Trafton G (eds) Proceedings of the twenty-ninth annual conference of the cognitive science society. Erlbaum, Mahwah, NJ, pp 503–508
Paige JM, Simon HA (1966) Cognitive processes in solving algebra word problems. In: Kleinmuntz B (ed) Problem solving: research, method and theory. Wiley, New York, pp 51–119
Palmer S (2003) Perceptual organization and grouping. In: Kimchi R, Behrman M, Olson CR (eds) Perceptual organization in vision: behavioral and neural perspectives. Erlbaum, Mahwah, NJ, pp 3–43
Patrick GTW, Gilbert JA (1896) On the effects of loss of sleep. Psychol Rev 3:469–483
Ranney M (1998) "Who are *you* to play God? Puzzles from an interloping 'rederminist' (reductive determinist)." Paper presented at the annual meeting of the Cognitive Science Society, Madison, WI
Ranney M, Cheng F, Garcia de Osuna J, Nelson J (2001) Numerically driven inferencing: a new paradigm for examining judgments, decisions, and policies involving base rates. Paper presented at the annual meeting of the Society for Judgment and Decision Making, Orlando, FL

Ranney MA, Rinne L, Munnich E, Yarnall L, Johnson T, Schank P (2007) "The 'Numbers, News, and Evidence' journalism curriculum: Might it boost everyone's reasoning?" Paper presented at the Learning and Teaching Workshop, Graduate School of Business, University of Chicago

Ranney M, Schank P (1995) Protocol modeling, textual analysis, the bifurcation/bootstrapping method, and Convince Me: computer-based techniques for studying beliefs and their revision. Behav Res Methods Instrum Comput 27:239–243

Ranney M, Schank P (1998) Toward an integration of the social and the scientific: observing, modeling, and promoting the explanatory coherence of reasoning. In: Read S, Miller L (eds) Connectionist models of social reasoning and social behavior. Lawrence Erlbaum, Mahwah, NJ, pp 245–274

Ranney M, Schank P, Hoadley C, Neff J (1996) I know one when I see one: How (much) do hypotheses differ from evidence? In: Fidel R, Kwasnik BH, Beghtol C, Smith PJ (eds) Advances in classification research, vol 5 (ASIS Monograph Series). Learned Information, Medford, NJ, pp 141–158, etc

Richman HB, Staszewski JJ, Simon HA (1995) Simulation of expert memory using EPAM IV. Psychol Rev 102:305–330

Rinne L, Ranney M, Lurie N (2006) Estimation as a catalyst for numeracy: micro-interventions that increase the use of numerical information in decision-making. In: Barab SA, Hay KE, Hickey DT (eds) Proceedings of the seventh international conference of the learning sciences. Lawrence Erlbaum, Mahwah, NJ, pp 571–577

Rosekind MR (2005) Managing work schedules: an alertness and safety perspective. In: Kryger MH, Roth T, Dement WC (eds) Principles and practice of sleep medicine, 4th edn. Elsevier, Philadelphia, pp 680–690

Sahakian WS (1968) History of psychology: a source book in systematic psychology. Peacock, Itasca, IL

Tobin KG (1987) The role of wait time in higher cognitive level learning. Rev Educ Res 57:69–95

Tweney RD (2003) Whatever happened to the brass and glass? The rise of statistical "instruments" in psychology, 1900–1950. In: Baker D (ed) Thick description and fine texture: archival research in the history of psychology. University of Akron Press, Akron, OH, pp 123–142, 200–205 (notes)

Tweney RD, Doherty ME, Mynatt CR (eds) (1981) On scientific thinking. Columbia University Press, New York

Wertheimer M (1987) A brief history of psychology. Holt, Rinehart, & Winston, New York

Author Biography

Michael Ranney's primary affiliation is in the Graduate School of Education at the University of California at Berkeley, where he chiefly serves programs in its Cognition and Development area. Prof. Ranney recently chaired SESAME, the interdepartmental Graduate Group in Science and Mathematics Education, and Ranney was Director of Berkeley's TEA$(CH)_2$EM physical science credential program. He is also an affiliated professor in Berkeley's department of psychology and he is on the executive committee of Berkeley's Institute of Cognitive and Brain Sciences—whose Cognitive Science programs he additionally serves as a faculty member. He was a Postdoctoral Fellow of Princeton University's Cognitive Science Laboratory, a Regents' Junior Faculty Fellow of the University of California, and a Spencer Fellow of the Spencer Foundation and the National Academy of Education. Ranney received MS and PhD degrees in cognitive/experimental psychology (University of Pittsburgh), and completed BA majors in psychology and molecular/cellular/developmental biology (University of Colorado, Boulder). His publications have engaged many fields, including philosophy, animal learning, and intelligent tutoring systems—as well as materials science and applied physics. Ranney heads the Reasoning Research Group, and he teaches a variety of courses on higher cognition, problem solving, and cognitive science. His research involves inference and the relative coherence of explanations, in both formal and informal domains; most of these domains involve elements of scientific and/or mathematical thinking, such as the realms of physics, biology, chemistry, medicine, the environment, and immigration policy.