Adv in Health Sci Educ (2016) 21:251–256
DOI 10.1007/s10459-016-9671-z

EDITORIAL

A bridge too far

Geoff Norman[1]

Published online: 21 March 2016
© Springer Science+Business Media Dordrecht 2016

In some of my darker moments, I can persuade myself that all assertions in education (a) derive from no evidence whatsoever (adult learning theory), (b) proceed despite contrary evidence (learning styles, self-assessment skills), or (c) go far beyond what evidence exists. I suspect most readers of AHSE are aware of the first two kinds of assertion, but in this editorial I want to elaborate on the third, the challenge of arriving at general conclusions about the way the world works based on the empirical evidence derived from limited studies.

It is not a new idea; like many of the things I think about, I can trace its roots back a few decades. But it has come to the fore as a result of a couple of recent occurrences that have shaken my faith in some research that I thought was impeccable. "Faith" may seem a strange term, but in this case, it feels apt. I really have gone from believer to skeptic as a consequence of some recent evidence that has come to light.

I am speaking of the recent research in the "Science of Learning" paradigm. In the past few years, many iconic figures in cognitive psychology have moved over to medical education and reported studies based in cognitive theories of learning. The studies are elegant; the theories are robust and time-tested. And the findings are truly impressive—as long as you don't look too close. Based on understanding of the way people learn, they demonstrate several simple but incredibly effective experimental manipulations that can have a powerful positive effect on learning. These are: (1) Interleaved practice—mixing examples from several categories together so that, to solve problems, you have to actively identify the features that distinguish one category from another, (2) Distributed practice—spacing learning sessions over time leads to reinforcement and better learning, (3) Test-enhanced learning—instead of simply studying the material, using in small tests that repeatedly revisit the content.

✉ Geoff Norman
norman@mcmaster.ca

[1] McMaster University, Hamilton, ON, Canada

One problem with these three interventions is that they are not nearly as universal as may seem. All focus on strategies to make practice more effective. They provide no guidance about strategies that may facilitate initial learning. To state the obvious, practice in problem solving is only useful to the extent that the end goal of learning is to solve problems. It is hard to imagine what mixed practice would contribute to a course on existential philosophy, quantum mechanics, or Shakespeare. Test enhanced learning may help the student learn some relevant facts, but that is hopefully not the primary goal of most courses (we'll return to this in due course).

But there is a further big disappointment that lies in the fine print. In order to test these theories, researchers have devised materials that really exemplify the kinds of skills where the strategy would be effective. For example, much of what we know about mixed practice derives from distinguishing classes of butterflies and Impressionist painters (also, incidentally, from many studies in motor learning that predate the resurgence of cognitive studies by decades). Now these may sit comfortably with biologists and art historians (although I expect both would proclaim there is far more to their discipline than simply telling examples apart). And we can find analogous areas in medicine that fit well with this paradigm such as reading ECGs or distinguishing heart sounds. But the point is that there is no attempt to identify the specific characteristics that make a set of materials amenable to such manipulations. And more egregious, there is little attempt, when the studies are published, to systematically explore the limits of generalizability of the findings—the boundary conditions set by the materials. Authors do not say, "If you are teaching students to identify a bunch of confusable categories in visual materials that can be displayed and learned quickly, try interleaved practice". They just say, "Interleaved practice works."

In particular, it is now recognized (but not by authors of the original studies) that test-enhanced learning—using mini tests repeatedly—is effective for recall of isolated and unrelated facts, but is relatively ineffective when any kind of transfer—even simply rewording the question or changing the distractors (Van Gog and Sweller 2015; Agarwal et al. 2012), although there are some exceptions (Larsen et al. 2013).

Similar constraints on the generalizability of mixed practice have emerged recently. On the one hand, studies of cognitive category learning using conceptually complex materials like ECGs have shown that mixed practice is only effective after some level of mastery has been attained using blocked practice. And motor learning, where it all began, now reveals that while mixed practice is good for simple actions, it has no advantage for more complex activities (Ranganathan and Newell 2010).

Unfortunately, it is not just this field that is apparently unaware of the limits imposed on their generalizations by the choice of materials. As another example, the critical role of context in learning is viewed as a precondition for instructional strategies like workplace-based learning, situated cognition, etc. Inevitably, the evidence for their claims includes the classic study of Godden and Baddeley with the Cambridge University Diving Club memorizing lists of unrelated words underwater and on land. No one seems to notice that, if you're trying to learn 36 unrelated words, you may well grasp at any crutch. When the study was repeated using medical materials, no effect was found (Koens et al. 2003).

The vast literature on deliberate practice is based on the assertion that the single determinant of expertise is practice—deliberate, structured practice with feedback (Godden and Baddeley 1975; Ericsson et al. 1993; Ericsson 2004). But there are some chinks in that armour. First, while deliberate practice does have a critical role to play in some areas of expertise, notably chess and music, it is not the sole determinant of success that popular treatises (Gladwell 2008) promise. There is in fact enormous variation in time to master even in well studied areas like chess (Gobet and Campitelli 2003). And while some studies

have indicated that general aptitude is not a significant predictor of performance, other studies contradict this. Finally, to revisit a common theme, deliberate practice is not a good predictor of expertise in more complex and multifaceted domains like the professions (Kulasegaram et al. 2013).

One more example, then we'll turn from observation to explanation. An area of medical education that has interested me over my entire career is clinical reasoning. Recently the field has been dominated by a concern with diagnostic error, which in turn is almost universally blamed on cognitive biases. A major protagonist in this perspective is Pat Croskerry, who has written extensively about dual processing models of expertise and the central role of cognitive biases in so-called "System 1" reasoning (Hambrick and Engle 2002; Croskerry 2003). However the central role of cognitive bias is a recurrent theme as far back as 1980s.

When you try to track down the origins of the theory, it emerges that there are very few studies attempting to demonstrate cognitive bias as a determinant of diagnostic error. Moreover, what few studies are available are based on either an experimental manipulation designed to create a particular bias such as availability (Mamede et al. 2010), or on a retrospective review (Graber et al. 2005) which is itself vulnerable to hindsight bias. Instead, a common strategy in the many armchair writings about cognitive bias in medicine is to cite the extensive research program in the 1970s and 1980s conducted by Tversky and Kahneman (1974). What is forgotten is that the virtually universal characteristic of these studies is that they were conducted on first year undergraduate psychology students using questions of dubious relevance (e.g. Does the letter "R" occur more often in the first or third position of a word?").[1]

The larger question, however, is precisely what is the relevance of this research to our understanding of diagnostic expertise. It reveals nothing about expertise, since this was not examined in their studies. In fact the very few studies that have looked at bias and expertise show that generally, experts are less vulnerable to bias than novices. Nor does it provide insight into possible interventions to mitigate bias, since this was not part of their research program. Indeed, Kahneman (2011) is adamant that cognitive biases (a) originate entirely in System 1, (b) Are hard-wired and irremediable, and (c) are unrelated to expertise. Why would he think otherwise? He has no data to prove otherwise.

So, while some of the researchers in these areas have been quick to proclaim the superiority of evidence and lament the extent to which educators fall prey to seductive "theories":

> The field of education seems particularly susceptible to the allure of plausible but untested ideas and fads (especially ones that are lucrative for their inventors). One could write an interesting history of ideas based on either plausible theory or somewhat flimsy research that have come and gone over the years. And.…once an idea takes hold, it is hard to root out. (Roediger and Pyc 2012)

It seems to me that the pot calls the kettle black. While it is true that research in this tradition is often based on elegant experiments with impressive findings, far too frequently, the particular materials have specific characteristics that are chosen to exemplify the phenomenon under study, and this in turn seriously constrains the generalizations possible from the study findings.

[1] The answer is "third". But most people say "first", because they can more easily recall "**r**emember" than "un**r**ealistic". However, as Lopes (1991) pointed out, the coin is rigged. Twelve of 20 consonants arise more often in the first position but Kahneman and Tversky picked one of the minority.

A disclaimer: Like most of my ideas, this is is not new, only rediscovered, in a different time and place. It has been formulated in research methods courses as a contrast between "internal" validity—the extent to which the study findings are believable, which is where the usual methodology stuff like randomization, confounders, statistical power etc. play out, and "external validity"—the extent to which the findings can be generalized to other situations, such as the "real world"—(whatever that stands for). It's important to recognize from the outset that this is not the "validity" of Messick (1989), Downing (2003) and Kane (2001), which refers to the validity of a measurement tool. This is far broader, and challenges the generalizability of any research study using any design.

And a second disclaimer: I can detect a certain smugness from qualitative researchers, who will remind me that they have been saying for years, nay decades, (Guba and Lincoln 1994) that your can never generalize beyond the conditions of a particular study. But it seems to me that this rejects the baby with the bathwater. We're not talking about not being able to generalize; we're talking about how far you can generalize. At the risk of sounding like an unrepentant positivist, it seems likely to me that some things like neutrons do generalize over the entire distance of the universe and over time from the origins of the universe. Maybe the first few milliseconds after the big bang, things were different, but I think that neutrons haven't changed much in the last 13.5 billion years. On the other hand, some things do not generalize. The challenge is to figure out which is which and how much we can generalize.

The issue of external validity rears its head in many quarters. Clinicians bemoan the findings of randomized trials, which are typically derived from highly atypical populations, saying things like "But that doesn't apply to my patients!". Science of learning folks and other psychologists are not blissfully unaware of the problem of generalization from the controlled lab study, but they worry more about the tendency to examine short term effects not long term learning in lab settings not classrooms, and worry less about the materials they create to test the effects.

What does all this have to do with the title of the editorial, "a bridge too far"? The whole problem was elegantly framed by Cornfield and Tukey (1956), in a formulation called the "Cornfield–Tukey bridge argument", which I described in a previous editorial in a discussion of psychometric validity (Norman 2015). In brief, they imagine a river with an island in the middle, which has the special property that it can move. The goal is to generalize from the study findings—the near bank—to a general assertion—the far bank. The distance from the near bank to the island represents internal validity—generalizing to other identical situations, and from the island to the far bank is external validity. And the basic idea, which captures all of the examples above, is that as we exert more and more control over the study, to increase internal validity, we move the island closer to the near bank and sacrifice external validity.

Medical education does have some very nice features that help us avoid this trap. Medical students, our typical participants, will show their impatience if the have to learn irrelevant or fictitious categories. So to some degree, safeguards are built in. Still, every time I embark on another study using written case protocols, there is just a bit of me that looks over my shoulder to see the ghost of John Tukey staring down and wagging a finger. Fortunately, awareness of the issue has led to some research demonstrating nicely that in several domains, low fidelity (i.e. unrealistic) simulations with the critical elements correctly portrayed do result in transfer of learning (Durning et al. 2012; Norman et al. 2012). But we should be constantly vigilant about the constraints of empirical research and suspicious of grand claims.

One final point: I am not challenging the conduct of lab-based research. No one likes carefully controlled experiments better than I; take them away and 2/3 of my c–v vanishes. What I am challenging is the nature of the inferences made from the research. As Mook (1993) has described, a research study need not be authentic, real-world, or high fidelity to be valuable. The value rests, however, with the nature of the generalizations made from the findings. I fear that, in the examples I have described, the "take home" messages go far beyond the evidence. And it's the take-home messages that are taken home.

References

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review, 24*, 437–448.

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *The Annals of Mathematical Statistics*, 907–949.

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine, 78*(8), 775–780.

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education, 37*(9), 830–837.

Durning, S. J., LaRochelle, J., Pangaro, L., Artino, A. R, Jr, Boulet, J., van der Vleuten, C., & Schuwirth, L. (2012). Does the authenticity of preclinical teaching format affect subsequent clinical clerkship outcomes? A prospective randomized crossover trial. *Teaching and Learning in Medicine, 24*(2), 177–182.

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine, 79*, S70–S81.

Ericsson, K. A., Krampe, R., & Tesch-Romer, T. H. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*, 363–406.

Gladwell, M. (2008). *Outliers: The story of success*. Paris: Hachette UK.

Gobet, F., & Campitelli, G. (2003). The role of domain-specific practice, handedness and starting age in chess. *Developmental Psychology, 43*, 159–172.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology, 66*(3), 325–331.

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. *Archives of Internal Medicine, 165*, 1493–1499.

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. *Handbook of Qualitative Research, 2*(163–194), 105.

Hambrick, D. Z., & Engle, R. W. (2002). Effects of domain knowledge, working memory capacity and age on cognitive performance: An investigation of the knowledge-is-power hypothesis. *Cognitive Psychology, 44*, 339–387.

Kahneman, D. (2011). *Thinking, fast and slow*. London: Macmillan.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342.

Koens, F., Ten Cate, O. T. J., & Custers, E. J. (2003). Context-dependent memory in a meaningful environment for medical education: In the classroom and at the bedside. *Advances in Health Sciences Education, 8*(2), 155–165.

Kulasegaram, K. M., Grierson, L. E., & Norman, G. R. (2013). The roles of deliberate practice and innate ability in developing expertise: Evidence and implications. *Medical Education, 47*(10), 979–989.

Larsen, D. P., Butler, A. C., & Roediger, H. L, I. I. I. (2013). Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Medical Education, 47*(7), 674–682.

Lopes, L. L. (1991). The rhetoric of irrationality. *Theory and Psychology, 1*(1), 65–82.

Mamede, S., van Gog, T., van den Berge, K., Rikers, R. M., van Saase, J. L., van Guldener, C., & Schmidt, H. G. (2010). Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. *Journal of the American Medical Asssociation, 2010*(304), 1198–1203.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.

Mook, D. G. (1993). In defense of invalidity. *American Psychologist*, 379–387.

Norman, G. (2015). The negative consequencesof consequential validity. *Advances in Health Sciences Education, 20*, 575–579.
Norman, G., Dore, K., & Grierson, L. (2012). The minimal relationship between simulation fidelity and transfer of learning. *Medical Education, 46*(7), 636–647.
Ranganathan, R., & Newell, K. M. (2010). Emergent flexibility in motor learning. *Experimental Brain Research, 202*, 755–764.
Roediger, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition, 1*(4), 242–248.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.
Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review, 27*, 247–262.