Adv in Health Sci Educ (2011) 16:287–289
DOI 10.1007/s10459-011-9310-7

EDITORIAL

Now you see it, now you don't?

Geoff Norman

Received: 15 June 2011 / Accepted: 15 June 2011 / Published online: 5 July 2011

Over a period of a few weeks in the spring, the same article from the New Yorker (Lehrer 2010) was sent to me by three different people. The article describes an interesting phenomenon, in which all sorts of scientific discoveries, ranging from a class of highly successful antipsychotic drugs to a psychological effect called "verbal overshadowing" initially show strong effects, but over time, with repeated observations, the effects drift downwards to the point that they no are no longer significant.

At one level, this is actually fairly commonplace. Living in an epidemiology department, I've seen many treatments come and go. Some examples:

> Ulcers are caused by spicy food and treated with milk; ulcers are caused by stress and treated with tranquilizers; ulcers are caused by excess acid and treated with antacids; ulcers are caused by acid reflux and treated with H2 antagonists; ulcers are caused by H Pylori and treated with antibiotics. Ulcers are best treated with proton pump inhibitors.
>
> Vitamin E prevents heart attacks. Oop, no it doesn't.
>
> Mammography/PSA saves lives. Hmm… maybe not.
>
> Estrogen replacement in women prevents heart attacks. Oops, it causes them.

In fact, a Greek epidemiologist, John Ioannidis, has a longstanding research program studying the history of clinical research findings that "flip". In one study Ioannadis (2005) looked at 45 highly cited positive (>1,000 citations) clinical studies (for example, estrogen replacement and heart disease, aspirin and heart disease) and found that in 16%, subsequent studies reversed the finding, 16% found larger effects on subsequent study, 24% were not replicated, and less than half (44%) had followup studies that were consistent.

One aspect is research design; 5/6 non-randomized study findings were changed subsequently. However there are other explanations that have nothing to do with methodology. Some of it is simply the alpha error problem; retrospective analyses are inevitably going to turn up some associations, such as Vitamin E and cancer, or mobile phones and brain

G. Norman (✉)
McMaster University, MDCL 3519, 1200 Main St. W., Hamilton, ON L8N3Z5, Canada
e-mail: norman@mcmaster.ca

tumours, purely by chance, but such fishing expeditions rarely arise in clinical trials. Vested interests may have, in the past, tainted results such as the clinical trial of Vioxx (Bombardier et al. 2000), although this is less likely with greater oversight today. One review article showed that 16% of trials funded by non-profit agencies showed positive findings compared to 51% of drug-company funded trials (Als-Nielson et al. 2003). And one must keep in mind that many of the effects identified in clinical trials, although statistically significant, may be exceedingly small in an absolute sense, so can reverse quite easily. Effect sizes less than 0.1 are the norm (Lipsey and Wilson 1993), a point I will return to.

The larger question for us in education is, does the same thing happen here? Are there instances where a phenomenon is demonstrated, but then goes away. Regehr in his "Rocket Science" article (2010), describes, somewhat cynically in my view, a process whereby the literature can be biased in favour of positive results because researchers just keep messing around until something works than publish it (if at first you don't succeed, trial, trial, trial again). That process, if true, could certainly lead to unreproducible positive results. On the other hand, It seems to me that starting with the assumption that the study didn't work because of problems with your methods, materials, etc., and then trying another variation, is a pretty rational and suitably humble process. However, once you got that first positive result after a series of disappointments, you should properly view it as pure seredipity (and ignore the p value) until you're explored the alternatives, replicated and extended. Not everyone does that, I suppose, but neither does everyone just publish the one study of a dozen that worked.

But neither Regehr's anecdote nor mine constitutes evidence that the New Yorker phenomenon is present or absent in education. I don't think it is; certainly not as commonly as Ioannadis describes in the clinical world. I can think of very few examples where a phenomenon, based on credible evidence (which excludes learning style, adult learning theory, etc.) was eventually overturned. The seminal work by Eva and Regehr (2005) on self-assessment, continued in this issue of AHSE (Eva and Regehr 2011), as they remind us, builds on a large body of consistent evidence going back to Gordon's beautiful reviews published in 1991. The phenomenon of content specificity, as Norcini (2005) noted, is ubiquitous—a consistent finding in assessment for 30 years. Supervisor assessments have been shown to be unreliable since at least 1972 (Streiner 1985), and continue to be so (van der Zwet et al. 2011) although they remain almost universal in clinical education worldwide.

Admittedly, consistency is not always the case. I have recently come across one vivid counter-example. The idea that the context of learning has an impact on subsequent transfer is almost axiomatic in medical education, and has stimulated a number of learning theories, including problem-based learning, contextual learning, situated cognition, workplace-based learning (McGill et al. 2011), and cognitive apprenticeship, all of which argue that the context of learning should match the context of eventual application. This "authenticity" argument also underlies much of the rationale for the use of high fidelity (and expensive) simulators (McGaghie et al. 2010). Almost universally, articles espousing this view refer to a single study by Godden and Baddeley (1975), in which Cambridge University divers learned lists of words on land and underwater and then recalled the lists in a matched or unmatched environment. Unfortunately, however persuasive the idea, it has been very difficult to replicate it, both in basic psychology studies (Hockley 2008) and in at least one study in medical education (Koens et al. 2003).

Why the discrepancy between us and the clinical researchers? One possibility is we're simply dealing with larger and therefore more robust effects. Lipsey and Wilson (1993) "meta-analyzed" 302 meta analyses, encompassing 14,000 original studies and found that

the average effect size for educational and psychological interventions across all the meta-analyses was 0.50. It's a lot harder to make big effects disappear. In contrast to the clinical studies described above, they found no relation between randomization or not and effect size, or between study quality and effect size (as long as it had a control group). They did see evidence of publication bias. But none of these biases made the positive effects go away.

I point all this out because Lipsey also cites effect sizes for medical interventions. Their Table 6 shows that for many conventional therapies and mortality (e.g. CABG, aspirin for heart attack) the ES is much smaller, often less than 0.1. I Although we in education may not deal with life and death issues, we may be in a situation where our results are actually more credible and robust than our 'big brothers' (at least big in terms of cost per study) in clinical research. And as a result it may be that we suffer less from the "now you see it—now you don't" phenomenon that seems to plague medical research.

Maybe it's time we quit apologizing for our inadequacies.

References

Als-Nielson, B., Chen, W., Gluud, C., & Kjaergard, L. L. (2003). Association of funding and colculsions in randomized drug trials. *Journal of the American Medical Association, 290*, 921–928.

Bombardier, C., Laine, L., Reicin, A., Shapiro, D., Burgos-Vargas, R., Davis, B. et al. (2000). Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. *New England Journal of Medicine, 343*, 1520–1528.

Eva, K. W., & Regehr, G. (2005). Self assessment in the health professions: A reformulation and research agenda. *Academic Medicine, 80*, S46–S54.

Eva, K. W., & Regehr, G. (2011). Exploring the divergence between self-assessment and self-monitoring. *Advances in Health Sciences Education, 16*(3). doi:10.1007/s10459-010-9263-2.

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

Gordon, M. J. (1991). A review of the validity and accuracy pd self-assessments in health professions training. *Academic Medicine, 66*, 762–769.

Hockley, W. E. (2008). The effect of environmental context on recognition memory and claims of remembering. *Journal of Experimental Psychology: Learning, Memory and Cognition, 34*, 1412–1429.

Ioannadis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited research. *Journal of the American Medical Association, 294*, 218–228.

Koens, F., Ten Cate, O. T., & 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: Theory and Practice, 8*, 155–165.

Lehrer J. The truth wears off. New Yorker, Dec 13 2010.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. *American Psychologist, 48*, 1181–1209.

McGaghie, W. C., Issenberg, S. B., Petrusa, E. R., & Scalese, R. J. (2010). A critical review of simulatin-based medical education research: 203–2009. *Medical Educaiton, 44*, 50–63.

McGill, D. A., van der Vleuten, C. P. M., & Clarke, M. J. (2011). Supervisor assessment of clinical and professional competence of medical trainees: A reliability study using workplace data and a focused analytical literature review. *Advances in Health Sciences Education*, *16*. doi:10.1007/s10459-011-9296-1.

Norcini, J. J. (2005). Current perspectives in assessment: the assessment of performance at work. *Medical Education, 39*, 880–889.

Regehr, G. (2010). It's NOT rocket science. rethinking our metaphors for research in health professions education. *Medical Education, 44*, 31–39.

Streiner, D. L. (1985). Global rating scales. In: V. R. Neufeld & G. R. Norman (Eds.), *Assessing clinical competence*. Springer, New York, pp. 119–141.

van der Zwet, J., Zwietering, P. J., Teunissen, P. W., van der Vleuten, C. P. M, & Scherpbier, A. J. J. A. (2011). Workplace learning from a socio-cultural perspective: Creating developmental space during the general practice clerkship. *Advances in Health Sciences Education*, *16*. doi:10.1007/s10459-010-9268-x.