Adv in Health Sci Educ (2015) 20:299–303
DOI 10.1007/s10459-015-9598-9

EDITORIAL

Identifying the bad apples

Geoff Norman

Published online: 16 April 2015
© Springer Science+Business Media Dordrecht 2015

Thirty-five years ago, two social psychologists, Richard Nisbett and Lee Ross, wrote a classic book called "Human Inference: Strategies and Shortcomings of Social Judgment" (1980). In that book, they demonstrated how human judgments and actions are vulnerable to many contextual variables. One particular shortcoming they labeled the "vividness hypothesis"—A single vivid instance can influence social attitudes when pallid statistics of far greater evidential value do not' (p. 57).

Evidence of this psychological bias abounds. Politicians continue to garner votes claiming they are "tough on crime" despite the fact that crime rates have been steadily declining in all developed countries for 25 years; as one example of many, homicide rates in Canada are half what they were in 1999. How bad was 2014 for air crashes? Remember MH370 and MH 17? In fact, the number of civil aviation crashes was the lowest on record, and about 1/3 of what it was in 1970, despite a five-fold increase in passenger miles flown. What about all those people killed by jihadists? Violent death rates have been on a steady decline for millennia (Pinker 2011).

In a subsequent book, "The Person and the Situation" (2011), they cast doubt on the whole notion of stable personality traits as a predictor of how someone will behave in a particular situation, and show that the specifics of the situation are a far greeter determinant of behavior.

We'll return to the role of personality on some detail shortly. For the moment, let us briefly comment on the role of vivid examples in health professions education. And what does all this have to do with health professions education? Well, we have our own vivid examples. Perhaps the most egregious is Dr. Harold Shipman, a British GP who is estimated to have killed 250 of his patients and was eventually convicted of 15 murders. The publicity surrounding his trial and conviction led to calls to reform the educational process

G. Norman (✉)
McMaster University, Hamilton, ON, Canada
e-mail: norman@mcmaster.ca

so that such things do not happen again (Powis 2015). In particular, we have seen increased focus on "non-cognitive" factors, particularly professionalism.

In this issue, two papers are also focused on the issue of professionalism. In a review article, van Mook et al. (2014) examines the multiple issues related to the identification and remediation of "dyscompetent" residents, particularly in the area of professionalism. And an original study by Santen et al. (2014) extends the findings of two landmark studies by Papadakis et al. (2004, 2008), which showed that unprofessional behavior in practice was apparently predictable from performance in medical school. Both studies used a "case control" design in which individuals who had the outcome of interest (discipline by a state licensing body)—"cases" were compared to a random sample of doctors who were not reported—"controls". In the first study at UCSF, 68 cases were compared to 196 controls. Review of records as undergraduates showed that cases were twice as likely to have documented concerns or problems as controls (38 vs 19 %).

The Santen et al. (2014) study in this issue replicated and extends these findings by examining the routine student assessments arising from promotion committees, instead of creating a system geared to identifying professionalism issues.

Both studies use a case–control design. Case–control studies are frequently used when the outcome of interest, such as developing cancer or dying, is infrequent. This design certainly applies in the Papadakis study of 6330 graduates from UCSF over the time interval, 70 were disciplined by the California state board, a prevalence of about 1.1 %. And, of course, 6260 were not. And there's the rub. The disease we're screening for—documented unprofessionalism—has a very low prevalence. Under these circumstances, even very good diagnostic tests result in true positive cases that are swamped by false positives. Working through this example, if we used documented concerns as a medical student as a screening test to decide if a graduate should be allowed to proceed, we would detect 38 % of the bad apples or 27; but we would incorrectly label $.19 \times 6260 = 1190$ other graduates as unprofessional. The positive predictive value of the test is $27/(1190 + 27) = 2.2$ %. Similar data arise in the Santen study, where review of 20 years' data, involving over 2000 graduates, showed that 140 had poor performance in school, and only 29 were subsequently sanctioned by the state medical board.

So in the Papadakis study, for every 100 students who would have been denied graduation, if they had proceeded to implement a policy based on documented concerns, only two would end up reported to the State board. Similar statistics arise form the Santen study, as pointed out by the authors.

If you want to put an economic spin on it, if it costs $100,000/year to educate a doctor, that policy would result in a social cost of $400,000 \times 98 = $40 million of education costs based on the number of satisfactory students who had an unsatisfactory and then could not graduate, without even considering lost income in practice. (While the back-of-the-envelope economic analysis is mine, I can claim no originality for the epidemiologic analysis, which was first published by Colliver et al. (2007).

But if these unprofessional behaviours are longstanding, perhaps they are detectable at the time of admissions. This is the promise held out by Powis, who has argued repeatedly for the more widespread use of personality tests at admissions (2003, 2009, 2015). While advocating strongly and repeatedly for use of personality, he does admit (2015) that "The million dollar question {is} do non-cognitive tests predict better outcomes at medical school and beyond?" and answers it with the following:

> Notwithstanding the current dearth of predictive validity data, if it is accepted that the non-cognitive qualities listed above… are essential components of professionally

> competent doctors' skills, then it is surely acceptable to use such tests on face validity grounds (p. 258)

My own view is that it is surely *not* acceptable to make life changing decisions about a student's future using instruments that simply look good (6). But that is one opinion against another.

In fact, there is some useful evidence to inform this policy. Papadakis, in another published study (2007), looked at performance on a personality test (the California Psychological Inventory), using a subsample from her earlier study that had undergone the psychological testing as part of admissions. The sample was 19 cases (from the original 70) who had difficulty with the state board, and 26 controls (of 196 sampled from 6260), all of whom who had taken the CPI as part of admission to medical school. For the total score, the mean of the cases was 156 (SD = 14.7); for the controls 181 (SD = 11.7). This means that the case mean was 25/11.7 = 2.1 SDs below the control mean. So far so good.

Now let us imagine using these data for selection, by establishing a threshold score that students must attain to be considered for admission—a policy directly advocated by Powis (2015). We can look at the proportion of each group who are accepted or rejected, keeping in mind that our real denominator is 70 cases who will eventually get in trouble with the state board, and 6260 controls who won't.

If we were to set the threshold at 156, the "case" mean, then we'll miss 50 % of the cases, 35. And this is a z score of −2.1 for the controls, so we'll falsely label 2 % of the controls, 125, as unprofessional. And 81 % (125/(125 + 35)) of the people we have labeled would not have any problems in practice.

Clearly, a test that only detects 50 % of the cases is of little value. So let's rack it up to a sensitivity of 90 %; which is a Z value on the "Cases" distribution of 1.28. We will detect 63 of 70 cases. That means the threshold, in Z units on the "control" distribution is (−2.1 + 1.28) = −0.82, which equates to 21 % of the Control distribution below the threshold, or 1315. In short, similar to the previous calculation, 1315/(1315 + 63) = 95 % of the individuals identified by a low score on the psychological test would not have any further problems in practice. In the table below, I've computed the proportion of false positives (students refused admission who would NOT end up in trouble) for a particular true positive rate.

True positive rate	Number of cases identified	Number of controls labelled	Proportion cases
.5	35	125	.28
.6	42	188	.22
.7	49	376	.13
.8	56	626	.09
.9	63	1315	.05
.95	66.5	4257	.015

Clearly, any attempt to identify individuals who will be eventually subject to report to the State disciplinary board using personality tests comes at a serious cost in terms of denying access to many who would not have problems.

The underlying premise of this approach is that cognitive measures are inadequate to identify individuals who will become problems in practice. But is this necessarily the case?

Tamblyn et al. (2007) has studied the validity of the Medical Council of Canada Qualifying Examination in predicting complaints (quality of care and communication skills) to provincial licensing bodies. The MCCQE examination has two parts—a written examination, primarily multiple choice completed at graduation, and an OSCE completed 1 year later. In terms of predicting communication complaints, the relative risk of a complaint for a communication skill performance in the bottom quartile of the OSCE was 1.43; for the written exam score was 1.34. For quality of care complaints, the relative risks were 1.38 for communication skills and 1.54 for the written test. (Relative risks for the data gathering and problem solving parts of the OSCE ranged from .97 to 1.13, predicting nothing). So it appears that, however much they are disparaged, cognitive measures of performance are an important predictor of practice performance. The same conclusion came from a study by Teherani et al. (2005), who looked at postgraduate performance of residents as a predictor of disciplinary action in practice. Performance was measured two ways: by American Board of Internal Medicine in-training evaluations, and by the ABIM certification examination. Again, the hazard ratio in predicting discipline charges looked impressive—about 1.9. However, the ABIM certification examination was not far behind at 1.7. And as before, with a prevalence of disciplinary action of about 1 % in this sample, the results do not support the use of either measure as a "diagnostic test".

One other assumption pervades discussion of assessing "non-cognitive" or personality at admissions—the "either-or" hypothesis. It is presumed that the admissions committee must make a Faustian choice between selecting someone who is personable, professional and compassionate, or someone who is academically top-tier. Such a choice would only be necessary if there were a strong *negative* correlation between personal qualities and academic performance. But is there?

One study (Powis and Bristow 1997) showed a significant negative association between scores on a personal interview and high school grades. However, two more recent studies examined the relation between the MMI (a well-validated measure of non-cognitive skills) and university GPA. In the first study (Eva et al. 2004) the correlation was −0.21; in the second (Kulasegaram et al. 2010) the correlation was +.07. Wherever the true correlation lies, it would appear that there should be no problem identifying students who have *both* academic excellence and interpersonal skills. Moreover, when one examines the constructs measured by the current stte of the art personality test, the Neo-5 personality test, about the only consistent relationship with other measures that has emerged is a moderate *positive* relationship between conscientiousness and grades (Kulasegaram et al. 2010).

It is perfectly appropriate to devise admissions strategies, in-course performance indices, and certification procedures that include both academic and interpersonal measures. It is *not* appropriate to force a choice between one and the other. And it is folly to presume that we will ever be able to create an adequate diagnostic test to the ultimately rare disease of unprofessionalism.

References

Colliver, J. A., Markwell, S. J., Verhulst, S. J., & Robbs, R. S. (2007). The prognostic value of documented unprofessional behavior in medical school records for predicting and preventing subsequent medical board disciplinary action: The Papadakis studies revisited. *Teaching and Learning in Medicine*, *19*, 213–215.

Eva, K. W., Rosenfeld, J., Reiter, H. I., & Norman, G. R. (2004). An admissions OSCE: The multiple mini-interview. *Medical Education, 38*(3), 314–326.

Hodgson, C. S., Teherani, A., Gough, H. G., Bradley, P., & Papadakis, M. A. (2007). The relationship between measures of unprofessional behavior during medical school and indices on the California Psychological Inventory. *Academic Medicine, 82*(10), S4–S7.

Kulasegaram, K., Reiter, H. I., Wiesner, W., Hackett, R. D., & Norman, G. R. (2010). Non-association between Neo-5 personality tests and multiple mini-interview. *Advances in Health Sciences Education, 15*(3), 415–423.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliff: Prentice Hall.

Papadakis, M. A., Arnold, G. K., Blank, L. L., Holmboe, E. S., & Lipner, R. S. (2008). Performance during internal medicine residency training and subsequent disciplinary action by state licensing boards. *Annals of Internal Medicine, 148*(11), 869–876.

Papadakis, M. A., Hodgson, C. S., Teherani, A., & Kohatsu, N. D. (2004). Unprofessional behavior in medical school is associated with subsequent disciplinary action by a state medical board. *Academic Medicine, 79*(3), 244–249.

Pinker, S. (2011). *The better angels of our nature: The decline of violence in history and its causes*. Harmondsworth: Penguin.

Powis, D. A. (2003). Editorial selecting medical students. *Medical Education, 37*, 1064–1065.

Powis, D. A. (2009). Commentary. Personality testing in the context of selecting health professionals. *Medical Teacher, 31*, 1045–1046.

Powis, D. (2015). Selecting medical students: An unresolved challenge. *Medical Teacher, 37*, 252–260.

Powis, D. A., & Bristow, T. (1997). Top school marks don't necessarily make top medical students. *The Medical Journal of Australia, 166*(11), 613.

Ross, L., & Nisbett, R. E. (2011). *The person and the situation: Perspectives of social psychology*. London: Pinter & Martin Publishers.

Santen, S. A., Petrusa, E., & Gruppen, L. D. (2014). The relationship between promotions committees' identification of problem medical students and subsequent state medical board actions. *Advances in Health Sciences Education,*. doi:10.1007/s10459-014-9536-2.

Tamblyn, R., Abrahamowicz, M., Dauphinee, D., Wenghofer, E., Jacques, A., Klass, D., et al. (2007). Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *JAMA, 298*(9), 993–1001.

Teherani, A., Hodgson, C. S., Banach, M., & Papadakis, M. A. (2005). Domains of unprofessional behavior during medical school associated with future disciplinary action by a state medical board. *Academic Medicine, 80*(10), S17–S20.

van Mook, W. N. K. A., van Lujik, S. J., Zwietering, P. J., Southgate, L., Schuwirth, L. W. T., Scherpbier, A. J. J. A., & van der Vleuten, C. P. M. (2014). The threat of the dyscompetent resident: A plea to make the implicit more explicit. *Advances in Health Sciences Education,*. doi:10.1007/s10459-014-9526-4.