Difficulties in Recognizing One's Own Incompetence: Novice Physicians Who Are Unskilled and Unaware of It

BRIAN HODGES, GLENN REGEHR, and DAWN MARTIN

Kruger and Dunning[1] published an elegant series of studies illustrating that deficits in the ability to assess one's own competence were common among the subjects in the lowest-scoring quartile on a wide range of tasks such as recognition of humor and logical reasoning. These researchers demonstrated that subjects who had the highest scores underestimated their abilities but were able to recalibrate following exposure to the performances of others. Those scoring in the middle quartiles were generally accurate in assessing their skills, and this remained so after exposure to others' performances. People in the lowest quartile, however, greatly overestimated their abilities and failed almost entirely to correct their self-assessments following exposure to the performances of others. Kruger and Dunning concluded that those who know less also know less about what they know. In their study they raise the question of how to help learners who intrinsically overestimate their abilities and therefore do not perceive their own incompetence.

The work of Kruger and Dunning is intriguing for medical educators because, if replicated with doctors, it suggests that those with the least skill may be most at risk of inaccurately assessing their abilities. We set out to determine whether the same problems of self-assessment so consistently replicated in the studies of Kruger and Dunning could be demonstrated in family medicine residents. We reported in a previous study that the calibration of the self-assessment of novices improved when they viewed benchmark performances of others.[2] In that study, the self-assessments of residents remained less than perfect even after viewing others' performances. Having reviewed the work of Kruger and Dunning, we decided to see whether it was indeed the lowest performers in our group who were most inaccurate in their self-assessments. We also wondered whether those in the highest-performing group would underestimate their skills but recalibrate following exposure to the performance of others. With these questions in mind, we reanalyzed the study data following the method of Kruger and Dunning and, indeed, both of these findings were confirmed.

Method

Twenty-four first-year family medicine residents interviewed a standardized patient (SP) in a difficult "breaking bad news" scenario. Two expert faculty members with extensive experience in teaching communication skills observed the ten-minute interviews. Both the residents and the experts rated the residents' performances on six nine-point rating scales (knowledge, rapport, emotional control, flexibility, questioning skills, and overall). The residents were then shown four videotaped interviews of the same SP scenario in which interviewers demonstrated a range of ability from highly skilled to quite incompetent. The videos were designed to represent a spectrum of performances from incompetence to advanced competence. The four videos were chosen from an initial pool of ten videos that were evaluated independently by two communication experts and four academic family physicians. Reliabilities of the videos and the scales had been established and are described elsewhere.[2] Inter-rater reliability between the expert raters was .93 and Cronbach's alpha was .92, suggesting that it was reasonable to collapse the six rating scales into one score. It had also been shown in the previous research that first-year residents clearly performed differently than did more experienced residents, the former group showing statistically significantly lower initial correlation with expert observers' ratings. In an effort to replicate the work of Kruger and Dunning, only these novices were selected for the current analysis.

Following their own interviews, the residents were asked to assess the videotaped performances and were then given an opportunity to rescore their own performances. The residents' ratings of their own performances were then compared with scores of experts in two ways. First, the experts' ratings of the residents were sorted in tertiles and plotted against both the pre- and the post-video self-assessed ratings of the residents. Second, in an effort to standardize the use of scales across the residents, each resident's score was re-scaled relative to his or her mean assessment of the four benchmark videos. A z score was calculated for each resident's performance by subtracting the resident's mean score of the videos from his or her actual self-assessment and dividing by the standard deviation of the video ratings.

The use of z scores requires some explanation. The example of one resident illustrates how analysis of raw scores alone could have obscured the fact that some residents became significantly more accurate in their self-assessments following the videos. Resident 21 gave himself a slightly higher raw score than the experts did prior to his viewing the benchmark videos (6.0 versus 5.9, respectively). He rated himself still higher (6.9) after he had viewed the videos, making it appear that he was becoming increasingly *inaccurate* in his self-assessment. However, when his scores were analyzed relative to the way he rated the videos (z scores), the picture changed greatly. This resident rated all of the video performances much higher than did the experts (mean rating of 7.2 versus 5.6, respectively). Thus, the initial score this resident gave himself was actually lower than was his mean rating of the videos (6.0 versus 7.2). On the other hand, the experts rated his performance higher than their mean rating of the videos (5.9 versus 5.6). Thus, when this resident viewed the videos and raised his self-assessment, he actually got closer to the experts' opinion that he was better than average. Using the z scores, we see that the changes he made to his self-assessment made him more accurate and were not evidence of further overinflation, as analysis of the raw scores alone suggested.

Results

When the residents' raw scores were sorted by tertiles, those in the top and the bottom tertiles initially scored themselves inaccurately relative to the experts' scoring. This was true both for raw scores (Figure 1) and for scores expressed as z scores (Figure 2).

To examine whether residents in either group were changing their scores significantly in the directions of the experts' scores, we compared the pre- and post-video self-assessments for residents in both groups, using a two-tailed t test. Examining the raw scores alone, the changes were not significant ($p = .26$ for the low group, $p = 0.19$ for the high group). However, using the z scores rescaled relative to the videos, the change in self-assessment toward the experts' scores was only marginally significant for the low-performing group ($.16 \pm .28$, $p = .074$) but highly significant for the high-performing group ($.37 \pm .24$, $p = .002$). The difference in the

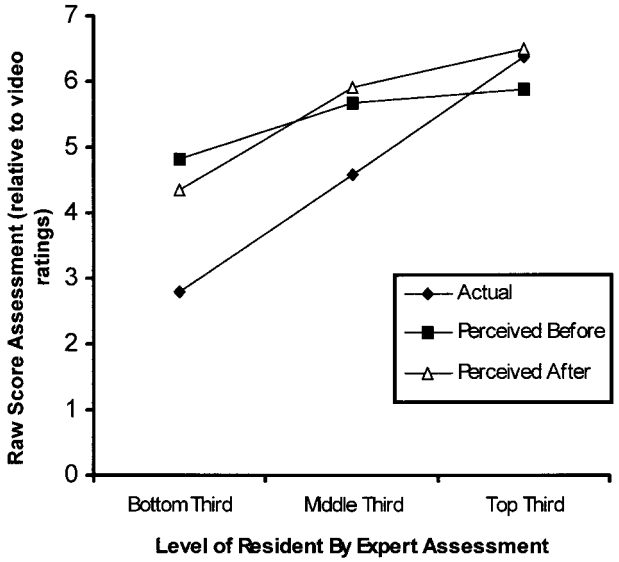

Figure 1. Raw score comparisons of expert assessments (actual) and resident self-assessments of performance (perceived) before and after seeing standards of comparison.

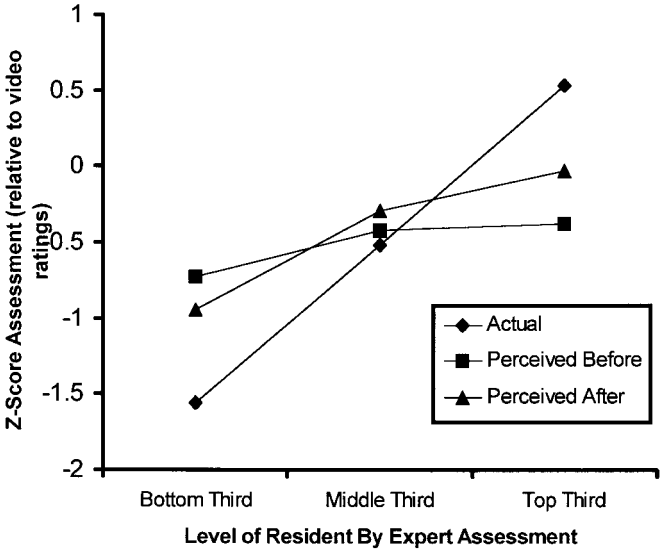

Figure 2. Z-score comparisons of expert assessments (actual) and resident self-assessments of performance (perceived) before and after seeing standards of comparison.

amount of correction between the two groups (.16 versus .37) was marginally significant (p = .06).

In fact, using the z score method, the difference in the patterns of correction between the two groups became quite evident. For the top-rated group, every single resident after viewing the videos demonstrated an appropriate but somewhat conservative correction in self-assessment scores toward the scores given by the experts. By contrast, the low-performing group demonstrated a variety of patterns after viewing the videos. Although three residents showed an appropriate shift toward the experts' scores, two residents adjusted too radically and began to underestimate their scores relative to the experts. One resident showed no change in self-assessment score at all, and two residents actually rated themselves higher after seeing the videos, despite the fact they they were clearly overestimating their performances prior to seeing the videos.

Discussion

Clearly this study was limited in scope, deriving results from a small sample of physicians from one specialty (family practice), in a specific domain (interviewing skills) in one setting. Thus, it is important to note that our findings may not generalize to physicians at other levels of experience or in other specialties. Furthermore, we used only one standardized patient case, and it is possible that this particular case might have had elements that made errors in self-assessment more likely. However, the results of this study do replicate the pattern that Kruger and Dunning found across a wide range of subjects in a wide range of tasks of knowledge and skills outside of medicine. While studies of self-assessment should be replicated in different contexts in medicine, even these preliminary results raise interesting questions for further research and for medical education.

We found that those residents in the highest performance group were able to recalibrate their self-assessments more accurately when presented with benchmark videos. Some residents in the lowest performance group were also able to do this, although not as consistently, and two of eight individuals actually worsened their already inflated self-assessments.

From a technical point of view, we feel this work might make an important contribution to the evolving area of self-assessment research. To date, almost all reports of accuracy of self-assessment have relied on comparisons of the raw scores of a group of subjects on a measure of performance, compared with scores of expert(s).[3] In this study, examining raw scores alone did not illustrate the problems of self-assessment shown in previous studies because different residents were using the measurement scale in different ways.[2] It was only by recalibrating their self-assessments relative to the way that they rated others that it was possible to see that the highest performers were gaining accuracy while the lowest performers were not. We would suggest that future studies of self-assessment consider using a similar method.

Turning to the educational implications, this work begins to reinforce a growing concern expressed by educators in health professional education.[3] Self-directed learning is based on the assumption that adult learners can identify and remedy deficits in their knowledge and skills. This is particularly important for self-regulating professions such as medicine, where continuing education is left entirely in the hands of individual professionals. It is only through accurate self-assessment that physicians can identify areas in which they are deficient in order to pursue further learning. But what if some physicians in some circumstances cannot accurately self-assess their skills? What if this inaccuracy persists, even when they are exposed to performances of peers? How will these learners overcome incompetence if left to direct their own learning?

Fortunately, improvements of self-assessment did occur in most of Kruger and Dunning's subjects and in our own. Indeed, we have confirmed that exposure to benchmark performances can lead to better self-assessments. However, that change in self-assessment was minimal or insignificant for the lowest group of residents, who were presumably at the greatest risk of incompetence.

Jeremy Taylor said: "It is impossible to make people understand their ignorance, for it requires knowledge to perceive it; and therefore, he that can perceive it hath it not." What are we to do about the bottom quartile or tertile? While more studies are needed to replicate and extend our findings, we feel that these results might suggest roles for any or all of:

- selection tests of self-assessment ability prior to medical training;
- teaching/testing self-assessment ability during medical school and residency;

- modeling of self-assessment and self-directed learning by teachers; and
- introduction of continuing education principles, including the development of self-assessment skills during undergraduate and postgraduate education.

This study was supported by a grant from Physician Services Incorporated Foundation.

Correspondence: Brian Hodges, MD, Toronto General Hospital, 8 EN212-200 Elizabeth Street, Toronto, Ontario, Canada M5G 2C4; e-mail: ⟨brian.hodges@utoronto.ca⟩.

References

1. Kruger J, Dunning D. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol. 1999; 77:1121–34.
2. Martin D, Regehr G, Hodges B. Improving self-assessment of family practice residents using benchmark videos. Acad Med. 1999;73:1201–6.
3. Gordon MJ. A review of the validity and accuracy of self-assessments in health professional training. Acad Med. 1991;66:762–9.