Practical Assessment, Research, and Evaluation

Volume 17 *Volume 17, 2012* Article 6

2012

# Indirect Measures In Evaluation: On Not Knowing What We Don't Know

Linda Heath

Adam DeHoek

Sara House Locatelli



## Recommended Citation

Heath, Linda; DeHoek, Adam; and Locatelli, Sara House (2012) "Indirect Measures In Evaluation: On Not Knowing What We Don't Know," *Practical Assessment, Research, and Evaluation*: Vol. 17 , Article 6.
DOI: https://doi.org/10.7275/00h8-7p49
Available at: https://scholarworks.umass.edu/pare/vol17/iss1/6

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.



Volume 17, Number 6, February 2012 ISSN 1531-7714

# Indirect Measures in Evaluation: On Not Knowing What We Don't Know

Linda Heath, Adam DeHoek, & Sara House Locatelli,
*Loyola University Chicago*

Evaluators frequently make use of indirect measures of participant learning or skill mastery, with participants either being asked if they have learned material or mastered a skill or being asked to indicate how confident they are that they know the material or can perform the task in question. Unfortunately, myriad research in social psychology has demonstrated that people are very poor judges of their own levels of accomplishment. In this paper, the social psychological dynamics that contribute to biased self-assessments are overviewed. These include the self-serving bias (e.g., Miller & Ross, 1975), the better-than-average effect (e.g., Alicke et al., 1995; Brown, 1986), and the overconfidence phenomenon (Kahneman & Tversky, 1979). Methods of correcting these biased reports are generally ineffective, as illustrated by Kruger and Dunning's (1999) findings that people lowest in mastery generally lack the metacognition even to understand what mastery looks like. As this type of person learns the skill in question, they often realize the level of their ignorance and lower their self-reported knowledge and skill levels. Although indirect measures of participant learning or mastery might tell us something about the level of confidence of the participants, they probably tell us little about actual ability or knowledge. Implications for applied research are discussed.

Many, if not most, interventions involve teaching someone something, whether it is teaching women in developing countries how to run their own businesses, teaching high school freshmen how to perform CPR, or even teaching teachers how to teach. This newly learned knowledge or skill will then presumably lead to more income and better living conditions, more lives saved, or better student learning. Linking the learning to the ultimate outcome, however, is not simple. For example, if students do poorly on tests, it might be that the teachers had not mastered the new skills or that the students were unmotivated or too stressed to learn. As plentiful as the complications are that intervene between the program and the ultimate outcome, however, evaluators realize that without actual knowledge or skill mastery in the first place, the intended downstream effects will never occur. Further, if evaluators cannot document that the knowledge or skill mastery took place, the staff of a failed program cannot know where the critical problems in the intervention reside, hindering further program refinement. Clearly, evaluating the extent of knowledge or skill mastery is critical not only for summative (or final assessment) evaluations, but also for formative (or developmental) evaluations.

## Use of indirect measures

Because direct evaluations of knowledge and skill mastery are not simple to conduct for practical and sometimes legal or ethical reasons, evaluators often make use of indirect, rather than direct, measures of learning. Such indirect measures are seen as useful by many teachers (Noonan & Duncan, 2005) and display consistent patterns among participants, though much weaker relationships to teacher assessments (Ross, 2006). Participants self-report how much they have learned, how confident they are they have mastered the skill, or how comfortable they feel with the skill or domain. This approach is based on the assumption that people can accurately judge their own abilities and knowledge-levels. Such indirect measures of knowledge and skill attainment are presumed to be a solution to the problems that direct measurement present. Unfortunately, such a move is going from the proverbial frying pan into the fire, as

indirect measures of learning are replete with their own problems.

This article overviews the factors that compromise the validity of indirect measures in applied settings. Research on these factors (e.g., Kruger & Dunning, 1999, 2002), as well as criticisms of such research (e.g., Krueger & Mueller, 2002) are published primarily in psychology journals, making these issues less well-known to evaluators and others who often use such indirect assessments in applied settings. The implications of this body of research for applied research are examined here. Among the many factors that can influence the validity of indirect measures of student learning, two are particularly problematic. First, most assessments of learning, but particularly those in areas that do not provide immediate, irrefutable feedback concerning success or failure, are prone to numerous distortions that are well-documented in psychology, such as the self-serving bias (e.g., Miller & Ross, 1975), the better-than-average effect (e.g., Alicke, Klotz, Breitenbecher, Yurak, & Vredenburg, 1995; Brown, 1986), and the overconfidence phenomenon (Tversky & Kahneman, 1979). Second, and perhaps the more serious problem, research by Kruger and Dunning (1999, 2002) has shown that people at the lowest levels of actual mastery lack the metacognitive understanding of what mastery even is, leading them to wildly over-estimate their own skills, often believing themselves to be above average. Each of these problems with indirect assessments of mastering knowledge and skills will be discussed in turn.

## Social Psychological explanations for biased self-evaluation in applied settings

Dunning, Heath, and Suls (2004) have described how and why these errors occur in areas ranging from health, to education, to the workplace. In that monograph they also document the lack of correspondence between people's self-assessments and the assessments of others, such as doctors or teachers, as well as the lack of congruence between people's self-assessments and their objective performance.

In the education area, accurate assessment of how much you have learned and what skills you really have would obviously be useful for determining how much to study, which advanced classes to take, and when to feel confident in performing a new task. Although the relationship between self-assessment of skills or knowledge and the objective level of that skill or knowledge is not zero, it is, in the words of Dunning, et al. (2004) "meager to modest" (p. 85). For example, people's estimates of their own IQ's correlate between .2 and .3 with their actual performance on IQ tests. Meta-analyses of the relationship between self-assessment and objective performance find correlations ranging from .21 to .39 (Dunning, et al., 2004). On the bright side, the correlation between students' assessments of their learning in classes and teachers' assessments of their learning does tend to increase from introductory to more advanced classes. Before becoming too excited, however, it must be noted that undergraduate pre-medical students' self-assessments' correlations with ratings of their own abilities and knowledge by teachers or supervisors actually got lower as time went on, with the students' final self-assessments of knowledge not correlating at all with their medical school board scores.

The major social psychological phenomenon at work in biased self-evaluation is the self-serving bias (e.g., Miller & Ross, 1975). This is the tendency for individuals to accept personal responsibility for success, while failing to take such responsibility for failure. This bias has been explained through both motivational and cognitive means. It is generally held that individuals have a desire for self-enhancement through both positive self-presentation and self-esteem maintenance. In times of success as well as in times of failure, individuals look for ways to present themselves positively to others. In cases of success, individuals may choose to publicly enhance their achievements, alerting others to the fact that they have succeeded; on the other hand, it is also possible that these individuals will downplay their achievements, if modesty is in order. Privately, people are likely to enhance their accomplishments, thus allowing themselves to maintain a high level of self-esteem. In cases of failure, individuals may publicly shun responsibility for an action, blaming others or the situation rather than themselves, if this course of action is likely to be perceived more positively by others. However, they may accept some of the blame for failure if humility will place them in a better light. In private, it is likely that individuals will blame the situation or others in order to maintain their own self-esteem. Thus, in both success and failure, individuals are able to present themselves positively. In addition, they are able to maintain the most positive self-esteem. In this way, people can experience self-enhancement both publicly and privately.

Just as the self-serving bias affects the responsibility that people take for a given outcome, it also affects the evaluation of one's own actions following both success and failure. Primarily, individuals are more likely to accept responsibility for expected outcomes rather than unexpected outcomes. Because individuals often intend

to and believe that they will succeed, people most often expect positive outcomes. In addition, people believe that they have more control over outcomes that they expect. Therefore, when individuals succeed, be it as students on an exam or investors on the stock market, they are likely to view this success as driven by their own efforts and talents, rather than merely the result of the situation. Research has shown that unexpected outcomes—be they positive or negative—will likely be attributed to external forces. As negative outcomes are often unexpected, these outcomes will more often be attributed to something outside of the individual and out of one's control. Therefore, when individuals fail, they will likely see this failure as the result of something that someone else did or did not do, or of the situation itself, rather than reflective of the self. Thus, using the same examples as those above, if students fail an exam, or brokers lose a significant amount of money, they are likely to perceive this as outside of their control. When it comes to self-evaluation, individuals are unlikely to recognize the true source of the outcomes in both successes and failures. Because of this, they are also unlikely to be able to effectively monitor how successful they would be at demonstrating the skills they believe they have mastered.

People also demonstrate self-serving effects and biased self-evaluation via a psychological phenomenon known as the better-than-average effect (e.g., Alicke et al., 1995; Brown, 1986). This phenomenon suggests that people believe that they possess positive traits to a greater degree and negative traits to a lesser degree than does the average person. Similarly, individuals believe that they know more and perform better on certain tasks than the average person. Therefore, if individuals rated their own abilities to complete specific tasks successfully—be it how effective they are at writing a paper or predicting changes in the stock market—they would likely state that they were better than the average person at such tasks. This effect is particularly pronounced when the comparison target is nonindividuated, such as "the average college student". In addition, College Board studies found that 60% of high school seniors rated themselves in the top 10% in terms of "ability to get along with others," and 25% rated themselves in the top 1% on this ability. The effect does not only hold for students. Ninety-four percent of college professors claim to do above-average work (Cross, 1977).

Finally, Tversky and Kahneman (1979) demonstrated an over-confidence phenomenon, whereby people's confidence in their assessments is non-significantly related to their accuracy. Clearly, we don't know what we know, but we are confident we do. This phenomenon exacerbates the biases discussed above. Not only are we wrong, but we are quite confident we are right!

## Attempts to increase the validity of biased self-evaluations

Unfortunately, remedying the problem of biased self-evaluation is not simply a matter of giving clearer feedback and more practice making these self-assessments for students to get it right. A study by Hacker, Bol, Horgan, and Rakow (2000) showed that higher performing students did get better able to predict their exam grades as the semester wore on, but poorly performing students continued to wildly over-estimate their performance, being not at all fazed by their repeated failures to achieve those high scores they had previously predicted.

So what's going on? Is it that students just don't know how to assess performance? No, because they do pretty well when assessing other students' performance. A meta-analysis found that the correlation between assessments of a peer and teachers' assessments averaged .72. So it seems to be something about assessing one's self that leads people astray.

One factor that contributes to erroneous self-assessments, as mentioned before, is that for many types of learning, there is not immediate, irrefutable feedback concerning mastery of a skill or knowledge area. For some learning outcomes (e.g., learning to do a triple-backflip, learning to insert an IV needle, being conversant in a foreign language with native speakers of that language), students can receive immediate, irrefutable feedback and clearly know if they have failed to master the skill. They fall on their heads, the patient screams in pain and no medicine flows through the IV, or the native speaker looks at the student in total incomprehension. An indirect measure of this type of learning outcome (e.g., "How confident are you in your ability to perform a triple-backflip?") might lead the student to ponder "Well, the last five times I tried it I landed on my head, so I guess I'm not very good at it." Nevertheless, motivational biases and selective recall could still lead to answers that are a bit more positive than the number of head-landings would warrant. This said, the measure could still be useable, given that its imperfections are recognized.

When we enter the realm of most student learning outcome assessments, however, we leave immediate,

irrefutable feedback behind. "How confident are you in your understanding of the causes of the Civil War;" "how confident are you in your ability to write a grammatically correct essay," and "how well do you understand the auditory system" all lack the immediate bump on the head or screaming patient of the previous examples. Students might factor in their course grades as proxies for their mastery, but that still leaves ample room for motivational biases, selective recall, and other psychological factors to distort their responses.

One clue that psychological factors are distorting self-assessments is the fact that self-assessments are overwhelmingly more positive than warranted. If people were just really bad at judging their own abilities and levels of knowledge, without a psychologically motivated component, we would expect their self-assessments to form a normal distribution around the actual ability level. Some people (or all people some of the time) would wildly over-estimate their abilities, and some people (or all people some of the time) would wildly under-estimate their abilities. Consequently, the average would hover pretty near the actual ability level. This is, in fact, the pattern we see when people are assessing peers' performance. But that pattern doesn't appear with self-assessments. Most people assess themselves as being "above average" on any given task, sort of the Lake Wobegon Effect. There is a rosy glow that surrounds our self-estimates. (To be thorough, it should be noted that very high-achieving people initially under-estimate their own percentile ranking on various skills. They fail to recognize not necessarily how good they are, but how bad most other people are. After feedback about their performance, however, high-performing people begin to give more accurate self-assessments (Kruger & Dunning, 1999).)

If the only problems with indirect measures of student learning were ones that resulted in an inflation of actual learning, these measures could be made useful by simply applying a correction factor, making program comparisons using the same distorted measure. So if Psychology majors reported mean confidence of 8 and English majors reported mean confidence of 6, if the only factors distorting the reports were creating main effects of rosy-glow, an administrator could conclude that the Psychology program was doing a better job than the English program at meeting learning outcomes, even though the actual achievement was probably lower than the 8 and the 6 would imply.

However, there is an issue underlying self-evaluation that makes applying this correction impossible. This is the discrepancy between knowledge and awareness of knowledge. While it is true that people are often aware of the things that they know, it is equally likely if not more common that they are unaware of what they do not know. Kruger and Dunning (1999) have discussed the topic of the interaction between knowledge and the awareness of knowledge. They claim that people who are deficient at a task are doubly disabled. Primarily, they do not know how to perform the task at hand. Secondly, although not of lesser importance, these individuals do not realize that they are not successful at completing the task. This phenomenon shows itself through indirect measures of learning. Individuals at the lowest levels of mastery inflate their abilities most and those at the highest levels of mastery actually understate their abilities. Kruger and Dunning (1999) showed this pattern of the least competent being least able to assess their own competence across judgments ranging from ability to judge funniness of jokes to grammatical skills to logic abilities. In all cases, participants whose abilities were objectively in the lowest quartile overrated their own abilities and often considered themselves to be above average. Kruger and Dunning argued that those in the lowest quartile lack the meta-cognitive ability to understand what "good performance" looks like, and are thus unable to recognize their own mistakes as failures. Kruger and Dunning showed that giving the students feedback about how well others had performed did not lower their ratings of their own abilities (although this feedback did cause students in the top-quartile to increase their ratings of their own ability more in line with their actual ability).

## Remedies

The only way Kruger and Dunning found to increase the accuracy of the ratings of students from the lowest quartile was to teach them the skill in question. Once they had been trained on what that skill actually involved, they reported that their own skill level was lower than they had reported it prior to such training. So, paradoxically, once participants had some clue about what the skill actually involved, they lowered their ratings of their own ability. Applying this finding, a professor who succeeds in teaching students the actual skills involved in the learning outcomes might be met with lower mean self-assessments of achievement from students, as the students in the lowest quartile come to report their ability more accurately.

As any professor knows who has heard students swear they know the material backward and forward, only to next hear them claim the retina is in the middle

ear, students often persist in believing they have mastered material even in the face of exam scores to the contrary. How many students have argued (to no avail) that they totally understand the material but the test didn't measure their knowledge correctly, or they know the material but they just couldn't produce it at the time of the test, or, nearer the end of the meeting with the professor, maybe they didn't know the material on the last test, but they certainly will know it now? If Kruger and Dunning are correct, these students don't even know what "knowing the material" is, let alone how to get there.

The realization that self-assessments generally do not accurately reflect actual learning or mastery leads to the inevitable conclusion that real program effects cannot be known with any certainty through indirect measures of learning or accomplishment. Although self-assessments might be useful for motivating students or measuring perceived learning, they do not measure actual learning. Real learning can only be assessed via direct measures of the skill or knowledge that was meant to be learned. Direct assessments are often difficult to do, time-consuming, expensive, and fraught with attrition and other practical problems. How do researchers convince teachers who are racing to cover all the material that time is needed for direct measures of learning? Who funds the long-term follow-up, years after the program (and likely the mandate that funded it) has ended? How can researchers gain cooperation from administrators who were not involved with the original study? Solutions to these problems need to be built into the original research design and contract. Efforts to educate funders and consumers of research about the limitations of indirect measures are necessary to ensure that direct measures are properly understood and valued.

In summary, to build evaluations on indirect measures of what participants report they have learned or mastered is to risk untold confusion and inaccurate conclusions. Programs that succeed in teaching the participants to assess their own performance accurately would fare very poorly in comparison with programs that let natural biases bloom, and programs that could reach the participants at the bottom of the ability distribution and teach them what mastery looks like would fare least well of all, as the participants might then accurately report that they actually have not mastered the material. Indirect measures based on what participants think they have learned or how confident they are in their abilities are not totally useless. For example, they can tell us if the students happily think they are wizards at statistical analysis or understanding the Civil War or writing essays. They just don't tell us if those confident students think the retina is in the ear. To borrow a term from the introduction to the Dunning, Heath, and Suls monograph, we won't know if the participants have actually learned or if they are just "blissfully incompetent."

## References

Alicke, M.D., Klotz, M.I., Breitenbecher, D.I., Yurak, T.J., & Vredenburg, D.S. (1995). Personal contact, individuation and the better-than-average effect. Journal of Personality and Social Psychology, 68, 804 – 825.

Brown, J.D. (1986). Evaluations of self and others: Self-enhancement biases in social judgments. Social Cognition, 4, 353-376.

Cross, P. (1977). Not can but will college teaching be improved? New Directions for Higher Education, 17, 1-15.

Dunning, D., Heath, C. & Suls, J. (2004). Flawed self-assessment: Implications for health, education, and the workplace. Psychological Science in the Public Interest, 5, 71 – 106.

Hacker, D.J., Bol, L., Horgan, D., & Rakow, E. (2000). Test prediction and performance in a classroom context. Journal of Educational Psychology, 92, 160 – 170.

Kahneman, D. & Tversky, A. (1979). Intuitive prediction: Biases and corrective procedures. Management Science, 12, 313-327.

Krueger, J. (1998). On the perception of social consensus. Advances in Experimental Social Psychology, 30, 163-240.

Krueger, J. & Mueller, R.A. (2002). Unskilled, unaware, or both? The contribution of social – perceptual skills and statistical regression to self-enhancement biases. Journal of Personality and Social Psychology, 82, 180-188.

Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. Journal of Personality and Social Psychology, 77, 221 - 232.

Kruger, J. & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77, 1121-1134.

Kruger, J. & Dunning, D. (2002). Unskilled and unaware—but why? A reply to Krueger and Mueller

(2002). Journal of Personality and Social Psychology, 82, 189 – 192.

Miller, D.T. & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? Psychological Bulletin, 82, 213-225.

Noonan, B. & Duncan, C.R. (2005). Peer and self-assessment in high schools. Practical Assessment, Research & Evaluation, 10 (17).

Ross, J.A. (2006). The reliability, validity, and utility of self-assessment. Practical Assessment, Research & Evaluation, 11(10).

## Citation:

Heath, Linda, DeHoek, Adam & Locatelli, Sara House (2012). Indirect Measures In Evaluation: On Not Knowing What We Don't Know. *Practical Assessment, Research & Evaluation*, 17(6). Available online: http://pareonline.net/getvn.asp?v=17&n=6

## Acknowledgements

An earlier version of this paper was presented at the 2005 meetings of the American Evaluation Association. Portions of this work were supported by funds from the Illinois Mathematics and Science Partnership program which was funded by NCLB, Title II, Part B, U.S. Department of Education in FY 2007, David Slavsky, P.I., Grant # 14-016-9000-01-51.

## Authors:

Linda Heath (Corresponding Author)
Department of Psychology
Loyola University Chicago
1032 W. Sheridan Rd.
Chicago, IL 60660
LHeath [at] luc.edu

Adam DeHoek
Department of Psychology
Loyola University Chicago
1032 W. Sheridan Rd.
Chicago, IL 60660

Sara House Locatelli
Department of Psychology
Loyola University Chicago
1032 W. Sheridan Rd.
Chicago, IL 60660