# UCLA

**Title**

Do formative assessments promote self assessment accuracy? A study of second year medical students' predictions about performance

**Permalink**

https://escholarship.org/uc/item/2gk7g47c

**Authors**

Relan, Anju
Guiton, Gretchen
Sopher, Michael
et al.

**Publication Date**

2006-08-01

# Do formative assessments promote self assessment accuracy? A study of second year medical students' predictions about performance

Anju Relan, David Geffen School of Medicine at UCLA, Gretchen Guiton, University of Colorado School of Medicine, Michael Sopher, David Geffen School of Medicine at UCLA, Josh Goldhaber, David Geffen School of Medicine at UCLA

Abstract: Self assessment accuracy is an essential precursor to the development of self-directed learning- a desired goal of medical education. The purpose of this study was to determine medical students' long term accuracy of performance prediction, evaluated in the context of formative assessments designed to provide practice for the summative assessment. Students were asked to predict their performance on the final exam over six formative assessments and at the end of the final exam. Students' confidence levels were predictive of near assessments, but were uncorrelated with final exam performance. Moreover, the ratings were stable over time. These findings are discussed in context of medical students' approach to learning in the second year of medical school.

**Introduction and Background**

The ability to diagnose one's strengths and weaknesses accurately and act upon these is critical to self-directed learning- an essential goal of contemporary medical education (Alexander, Kernohan, & McCullagh, 2004; Konje & Taylor, 1998; Mennin & Kalishman, 1998). Ideally, accurate self-assessments alert students to the gap between their own knowledge state versus the expected standard of skill or performance, based on which they choose to address these deficits via self-determined means, such as altering learning strategies, changing the focus of study or deliberately seeking help and collaboration (Fitzgerald, White, & Gruppen, 2003; MacDonald, Williams, & Rogers, 2003). Accurate self assessment has been recognized as a desirable skill among physicians striving towards self-directed learning (Ward, Gruppen, & Regehr, 2002). Fostering metacognitive awareness leading to accuracy in self-assessments must be an important objective of medical education to enable students to practice this skill with ease as physicians.

Before interventions on self-assessment skills training are designed, the existing level of this skill among students must be examined. Extant research on the comparison of estimated and actual performance of medical students has revealed some consistent findings, of which "illusion of knowing" is dominant (Ward et al., 2002). This is particularly true of low performing students, whose assessments tend to be inflated when compared to performance. It has also been hypothesized that self-assessment is a relatively stable quality, which may resist improvement with the development of expertise, or increase in knowledge. However, evidence to the contrary has emerged when self-assessments are gauged at more granular and differentiated levels, based on different types of tasks (Coutts & Rogers, 1999; Gruppen, White, Fitzgerald, Grum, & Woolliscroft, 2000; Woolliscroft, TenHaken, Smith, & Calhoun, 1993). In these studies, variables such as task characteristics may also intervene in the ability towards accurate self-assessments- with familiarity with the task over (Antonelli, 1997; Edwards, Kellner,

Sistrom, & Magyari, 2003; MacDonald et al., 2003).  From all these studies, it is apparent that self-assessment skills interact with a host of factors which mediate in its accuracy and consistency, suggesting that further research is needed to reveal trends.

The purpose of this study is to investigate second year medical students' self-assessment ability over a series of written formative assessments and a summative assessment.  It was hypothesized that extensive use of formative assessments, explicitly designed to simulate the final exam structure will lead to accurate self assessments for performance on the exam.  Previous studies have used multiple outcome measures to evaluate self-assessment, all of which have been in the context of summative evaluations.  The design of this study differs from these to the extent that the self-assessments were elicited on a series of formative assessments designed to provide students with practice on the summative assessment.  There was a short and long term component to assessments, with the expectation that formative assessments would cue students towards more accurate prediction on the summative assessment.

The research questions examined for the study were:

1. How precisely do students predict their performance on final exam within the context of formative assessments?
2.  How stable are students' assessments over the course administration with each implementation of formative assessment?

**Methods**

Subjects: Participants were 141 second-year students enrolled in UCLA medical school.  They were the second group of students to experience the implementation of an interdisciplinary, problem-oriented curriculum, in an 8-week course titled, "Foundations of Medicine: Cardiovascular, Renal and Respiratory Medicine."

Formative assessments:  Weekly, online formative assessments were implemented via ANGEL course management system (http://www.angellearning.com), the curriculum delivery tool used in the hybrid curriculum.  The purpose of the formative assessments was, a) focus students' attention on important topics in the Block; b) inform students of the types of questions to expect on the final exam; c) identify areas of weakness in instruction; d) provide immediate feedback on students' misconceptions; and e) offer students opportunities to practice critical thinking and to synthesize information from external sources.

The format of formative assessments consisted of timed, multiple choice questions.  These were made available to students every Friday, and were inaccessible by the following Monday morning.  Students could take assessment multiple times, however, only the score obtained during the first submission was recorded.  Formative assessments were a requirement in the course, but did not count towards the final grade. Most assessments consisted of approximately 20 questions.  Questions used were clinically

oriented, requiring students to identify or comment on the pathological origins of disease and examine diagnostic reasoning.

Summative assessment: A timed, summative assessment consisting of 76 items was administered in the ninth week of the course via ANGEL in a synchronous, proctored environment in a central computer lab. Items were randomized for additional security.

Self assessments:  Six of the seven formative assessments administered included one question on students' confidence in predicting their performance on the final exam.  The question was administered at the end, before students viewed their feedback on the assessment.  A 5-point Likert scale was used for responses (5=Very high; 4=High; 3=Modest; 2=Poor; 1=Very poor).  The question was: *How would you rate your level of confidence for passing FINAL EXAM questions related to this week's material?* The second question on predicting performance was included at the end of the final exam, before students viewed their score on the exam, *How would you rate your level of confidence for passing the present final exam?* The same Likert scale was used to elicit responses.

Data sources:  The data sources used were scores from formative assessments and the final exam, confidence assessment data gathered at the end of each formative assessment and at the end of the final exam.

Experimental Design: A correlational analysis was performed to analyze relationships between student confidence levels and performance on all assessments.  A repeated measures analysis was performed to determine the stability of student ratings over seven weeks, and of formative assessment performance over the same time period.  Students' responses on the survey items were correlated with their confidence assessments.  SPSS v. 13.0 was used for all analyses.

**Results**

The dataset was checked to satisfy assumptions of normality. The results are presented by research questions.

*1. How precisely do students predict their performance on final exam within the context of formative assessments?*

Table 1 shows the means, standard deviations and reliabilities (Cronbach's alpha) of formative and summative assessments.  The high mean and low standard deviation of summative assessment is striking, indicating the high effort expended on studying for the final exam.

Table 1: Means, standard deviations and reliabilities of formative and summative assessments.

| | *Mean* | *S.D.* | *Number of items* | *Reliability (Cronbach's alpha)* |
|---|---|---|---|---|
| Formative assessment Week 2 | 16.75 | 2.61 | 20 | .85 |
| Formative assessment Week 3 | 11.23 | 2.64 | 15 | .66 |
| Formative assessment Week 4 | 20.67 | 4.5 | 23 | .76 |
| Formative assessment Week 5 | 13.25 | 3.13 | 20 | .64 |
| Formative assessment Week 6 | 14.58 | 3.71 | 20 | .78 |
| Formative assessment Week 7 | 11.06 | 2.08 | 15 | .51 |
| *Summative assessment* | *67.84* | *7.25* | *76* | *.70* |

Table 2 shows the two-tailed results of Pearson correlation coefficients among means of formative assessment and final exam scores and the two confidence questions- at the end of the formative assessments and the final exam. The mean of confidence responses on all six formative assessments was used in this analysis.

The self assessment elicited after formative assessments was unrelated to the final exam, but was found to be significantly correlated with the formative assessment mean score. This result is surprising given that the question on self assessment was targeted towards achievement in final exam. This suggests students' ability for more accurate predictions over short term.

Table 2: Pearson correlation coefficients between the mean of all formative assessments, final exam mean, mean of self assessments across six formative assessments and the self assessment related to final exam.

| | Confidence in similar items on final exam | Final exam score | Mean formative assessment score | Confidence in final exam |
|---|---|---|---|---|
| Confidence in similar items on final exam | 1 | .165<br>.052 | .314(**)<br>.000 | -.311(**)<br>.000 |
| Final exam score | | 1 | .267 (**)<br>.001 | -.323(**)<br>.000 |
| Mean formative assessment score | | | 1 | -.277(**)<br>.001 |

** Correlation is significant at the 0.01 level (2-tailed).

Table 3 shows the correlations between students' predictions of performance on summative assessment with weekly formative assessment scores as well as corresponding content score in summative assessment. It also shows the correlations between performance on the weekly formative assessment and the items assessing that content on the final summative assessment.

Table 3: Correlations between weekly confidence assessment and performance on weekly exam and performance on matched content from final exam.

| Week | Correlations between weekly predictions and performance on weekly formative assessments | Correlations between weekly predictions and performance on weekly content items in summative assessment | Correlations between performance on weekly formative assessments and performance on related weekly content items in summative assessment |
|------|------|------|------|
| 2 | .373** | .162 | .045 |
| 3 | .213* | .016 | .264** |
| 4 | .225* | .090 | .191* |
| 5 | .049 | -.102 | .089 |
| 6 | .346** | .115 | .366** |
| 7 | .181* | .102 | .136 |

\* significant at the .05 level; ** significant at the .001 level.

These correlations suggest that students are moderately accurate in assessing their performance at the time of testing, but that assessment is not predictive of future performance. That only 3 of the 6 correlations between formative and summative assessment performance were significant suggests that the nature of students' performance itself may shift over time or that the items assessing performance changes.

Students' self assessment patterns relative to scores on formative and summative assessments over six weeks are plotted in Figure 1. The secondary y-axis is the likert scale for confidence assessments.
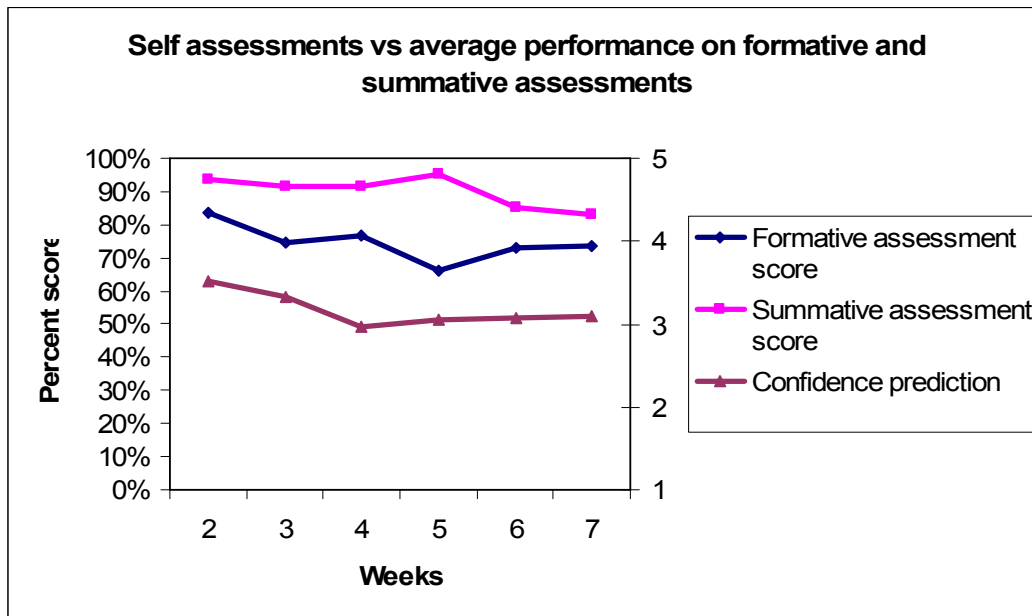
**Self assessments vs average performance on formative and summative assessments**

2. *How stable are students' assessments over the course administration with each implementation of formative assessment?*

A repeated measures analysis was performed with six confidence assessments presented after formative assessments as within-subjects factors. Mauchley's test of sphericity was found to be significant ($p<.00$), implying that the variances of differences among the confidence assessments were significant. Consequently, due to the violation of sphericity, Greenhouse-Geisser and Huynh-Feldt tests, which adjust for these differences were conducted. The F-ratio revealed no significant differences in any of these tests, indicating that there were no within-subject differences in confidence ratings across six self assessments (see Table 4). Thus students demonstrated consistent ratings in their performance prediction over a span of six weeks.

Table 4: Repeated measures analysis with self assessment ratings over six formative assessments as a within-subjects factor.

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig |
|---|---|---|---|---|---|---|
| Weekly assessments | Sphericity Assumed | .477 | 5 | .095 | .232 | .948 |
| | Greenhouse-Geisser | .477 | 4.249 | .112 | .232 | .929 |
| | Huynh-Feldt | .477 | 4.435 | .108 | .232 | .934 |
| | Lower-bound | .477 | 1.000 | .477 | .232 | .631 |
| Error(week) | Sphericity Assumed | 234.023 | 570 | .411 | | |
| | Greenhouse-Geisser | 234.023 | 484.440 | .483 | | |
| | Huynh-Feldt | 234.023 | 505.533 | .463 | | |
| | Lower-bound | 234.023 | 114.000 | 2.053 | | |

**Discussion**

In this study we sought to determine whether students' self assessments of their performance ability based on formative feedback would lead to accurate predictions for performance on the final exam. The task of formative assessments is to promote self monitoring of learning among students, so that they can effectively calibrate their time and effort into appropriate study behaviors. In a curriculum where students are extensively exposed to formative assessments, this weekly feedback constitutes an intervention in the curriculum to encourage self monitoring and self directed learning. Self assessments elicited in the context of formative assessments can provide a sense of students' ability to reflect on their present state of learning and make predictions about what is required to know and learn for the future.

A second question of interest was the stability of self assessment ratings over a period of six weeks, related to different topics covered. This line of research was undertaken to add to studies of medical students' self assessment where the nature of task and other variables have intervened to mediate such assessments (Antonelli, 1997; Coutts & Rogers, 1999; Edwards et al., 2003; Gruppen et al., 2000; MacDonald et al., 2003; Tousignant & DesMarchais, 2002; Woolliscroft et al., 1993).

We found that the near transfer of students' self assessments to accurately predict performance on formative assessments was stronger than far transfer on the final exam. However, given modest magnitude correlations, even this fell short of desired levels of calibration accuracy, given the low stakes nature of such assessments. Medical students tended to be cautious about their prediction of final exam performance- which may lead to higher anxiety and studying than is warranted. This finding needs to be interpreted along with that of the second question, demonstrating that students' patterns of assessment are stable, and perhaps based on more than one course. Alternatively, it is possible that students are using global motivational variables, such as self efficacy in estimating their performance, versus effectively evaluating current task needs.

Recent work by Eva, Norman, Regehr and Reiter (Eva, Cunnington, Reiter, Keane, & Norman, 2004; Eva & Regehr, 2005; Reiter, Eva, Hatala, & Norman, 2002) has questioned the value of current practices in self assessment research. The authors express concern that the very skills needed to achieve are what is needed for accurate self assessment, thus defying the purpose for eliciting such assessments from inexperienced professionals. Our research, however, is a precursor to a deeper understanding of the role of formative assessments in promoting greater self assessment accuracy and eventually, towards well calibrated, effective study behaviors. A new study is presently being undertaken to examine students' accuracy in predicting performance based on feedback on the formative assessments, time spent on self study each week and perception of difficulty of content. Findings from such research are essential in designing appropriate evidence based interventions designed to improve self assessment accuracy.

**References**

Alexander, S., Kernohan, G., & McCullagh, P. (2004). Self directed and lifelong learning. *Stud Health Technol Inform, 109*, 152-166.

Antonelli, M. (1997). Accuracy of second-year medical students' self-assessment of clinical skills. *Acad Med, 72*(10 Sup 1), S63-65.

Coutts, L., & Rogers, J. (1999). Predictors of student self-assessment accuracy during a clinical performance exam: comparisons between over-estimators and under-estimators of SP-evaluated performance. *Acad Med, 74*(10 Suppl), S128-130.

Edwards, R. K., Kellner, K. R., Sistrom, C. L., & Magyari, E. J. (2003). Medical student self-assessment of performance on an obstetrics and gynecology clerkship. *Am J Obstet Gynecol, 188*(4), 1078-1082.

Eva, K. W., Cunnington, J. P., Reiter, H. I., Keane, D. R., & Norman, G. R. (2004). How can I know what I don't know? Poor self assessment in a well-defined domain. *Adv Health Sci Educ Theory Pract, 9*(3), 211-224.

Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: a reformulation and research agenda. *Academic Medicine, 80*(10 Suppl), S46-54.

Fitzgerald, J. T., White, C. B., & Gruppen, L. D. (2003). A longitudinal study of self-assessment accuracy. *Med Educ, 37*(7), 645-649.

Gruppen, L. D., White, C., Fitzgerald, J. T., Grum, C. M., & Woolliscroft, J. O. (2000). Medical students' self-assessments and their allocations of learning time. *Acad Med, 75*(4), 374-379.

Konje, J. C., & Taylor, D. J. (1998). Formative assessment within structured training in obstetrics and gynaecology. *British Journal of Obstetrics and Gynaecology, 105*(2), 139-141.

MacDonald, J., Williams, R. G., & Rogers, D. A. (2003). Self-assessment in simulation-based surgical skills training. *Am J Surg, 185*(4), 319-322.

Mennin, S. P., & Kalishman, S. (1998). Student Assessment. *Academic Medicine, 73*(9), S46-S54.

Reiter, H. I., Eva, K. W., Hatala, R. M., & Norman, G. R. (2002). Self and peer assessment in tutorials: application of a relative-ranking model. *Academic Medicine, 77*(11), 1134-1139.

Tousignant, M., & DesMarchais, J. E. (2002). Accuracy of student self-assessment ability compared to their own performance in a problem-based learning medical program: a correlation study. *Adv Health Sci Educ Theory Pract, 7*(1), 19-27.

Ward, M., Gruppen, L., & Regehr, G. (2002). Measuring self-assessment: current state of the art. *Adv Health Sci Educ Theory Pract, 7*(1), 63-80.

Woolliscroft, J. O., TenHaken, J., Smith, J., & Calhoun, J. G. (1993). Medical students' clinical self-assessments: comparisons with external measures of performance and the students' self-assessments of overall performance and effort. *Acad Med, 68*(4), 285-294.