A longitudinal study of self-assessment accuracy

James T Fitzgerald, Casey B White & Larry D Gruppen

Aim Although studies have examined medical students' ability to self-assess their performance, there are few longitudinal studies that document the stability of self-assessment accuracy over time. This study compares actual and estimated examination performance for three classes during their first 3 years of medical school.

Methods Students assessed their performance on classroom examinations and objective structured clinical examination (OSCE) stations. Each self-assessment was then contrasted with their actual performance using idiographic (within-subject) methods to define three measures of self-assessment accuracy: bias (arithmetic differences of actual and estimated scores), deviation (absolute differences of actual and estimated scores), and covariation (correlation of actual and estimated scores). These measures were computed for four intervals over the course of 3 years. Multivariate analyses of variance and correlational analyses were used to evaluate the stability of these measures.

Results Self-assessment accuracy measures were relatively stable over the first 2 years of medical school with a decease occurring in the third year. However, the correlational analyses indicated that the stability of self-assessment accuracy was comparable to the stability of actual performance over this same period.

Conclusion The apparent decline in accuracy in the third year may reflect the transition from familiar classroom-based examinations to the substantially different clinical examination tasks of the third year OSCE. However, the stability of self-assessment accuracy compares favorably with the stability of actual performance over this period. These results suggest that self-assessment accuracy is a relatively stable individual characteristic that may be influenced by task familiarity.

Keywords clinical competence, *standards; education, medical, *standards, *methods; educational measurement, *standards; longitudinal studies; reproducibility of results; self-concept.

Medical Education 2003;37:645–649

Introduction

Accurate, career-long self-assessment of knowledge and skills is essential for physicians to maintain and improve their medical proficiency through self-directed education. Physicians who cannot accurately self-assess their knowledge and skills may be at greater risk for providing suboptimal care to patients.

The body of research on medical student self-assessment is less than would be expected, given the significance of this phenomenon.[1] However, existing studies suggest that there is a developmental component in medical students' ability to evaluate themselves and peers that lags behind their ability to perform specific skills.[2] As suggested by findings that the accuracy of students' self-assessment skills increases slightly over the course of education,[3] self-assessment ability may be modifiable by education. However, even if self-assessment is a learnable or modifiable skill, it appears likely that much of this learning has taken place in childhood and that by the time students enter medical school is largely fixed.[4] The limited evidence of improvement in self-assessment skills during medical education may reflect the relatively stable character of adult self-assessment or it may reflect the fact that students receive little practice in self-assessment.

Since 1995, we have conducted a series of self-assessment studies in which we established methods for measuring self-assessment using intraindividual analysis. Intraindividual analysis enables us to characterize the accuracy of individual students, as opposed to an interindividual analysis, which produce group-level estimates of accuracy. We have used these measures to address the analytical problems recently described by

Department of Medical Education, University of Michigan Medical School, Ann Arbor, Michigan, USA

Correspondence: James T. Fitzgerald, University of Michigan Medical School, Department of Medical Education, Towsley Center, Room 1200, Box 0201, Ann Arbor, Michigan 48109–0201, USA. Tel: +1 734 763–1153, Fax: +1 734 936–1641, E-mail: tfitz@umich.edu

Key learning points

Practising physicians need to assess their knowledge and skills accurately to maintain their medical proficiency through self-directed learning.

Medical student self-assessment accuracy appears to be influenced by task familiarity; the more familiar the task, the more accurate the self-assessment.

However, medical student self-assessment accuracy is reasonably stable over time and task when compared with the stability of actual performance, supporting the notion that self-assessment is a stable characteristic.

The results also demonstrate the value of an intraindividual methodology (as opposed to a group-level analysis) for studying self-assessment.

Ward *et al.*[5] and to understand better the components of medical student self-assessment.[6–8] Our studies indicate that self-assessment accuracy is not related to demographic (gender and ethnicity) or academic variables (academic performance and academic preparation).[9] Some of our preliminary investigations have also suggested that medical students' self-assessment abilities are stable over short periods of time[6] and over task.[10]

Our long-term goals have included acquiring a better understanding of self-assessment in order to help medical students grasp its importance to themselves and to their patients, to provide them with practice during medical school and to develop an intervention that might assist those with poor self-assessment abilities. To achieve these goals, it is critical to determine how stable self-assessment abilities are over time. Unless there is evidence that self-assessment accuracy is a relatively stable, consistent characteristic rather than a purely situational phenomenon, there is little point in considering educational interventions.

The focus of this study is to evaluate the temporal stability of medical students' self-assessment accuracy. By comparing three medical school classes' examination performance and self-estimates of this performance, we examined stability of medical student self-assessment accuracy from the first year through the third year of medical school.

Methods

The University of Michigan Medical School graduating classes of 1999 ($n = 163$), 2000 ($n = 169$) and 2001 ($n = 168$) were asked to provide estimates of their performance after completing each examination, quiz and lab examination in their M1 winter term, M2 autumn term and M2 winter term. For the class of 1999, 22 self-estimates were obtained for each student in the M1 winter term, eight in the M2 autumn term and 17 in the M2 winter term. For the class of 2000, 18 self-estimates were obtained for each student in the M1 winter term, 19 in the M2 autumn term and 16 in the M2 winter term. For the class of 2001, 18 self-estimates were obtained for each student in the M1 winter term, 18 in the M2 autumn term and 16 in the M2 winter term. During these terms, students were evaluated primarily on cognitive tasks, i.e. multiple-choice quizzes, labs and examinations. These self-estimates were provided on the same percentage correct scale used for quantifying their actual performance (0–100%).

At the end of the third year, after completing their required clinical rotations, students took a multiple-station objective-structured clinical exam (OSCE). They were asked to estimate their performance on each of the stations on a percentage correct scale. There were 10 stations for the class of 1999 and 13 stations for the two subsequent classes. The OSCE stations were primarily performance-based tasks, i.e. demonstrations of clinical skill, and differed from the classroom-based knowledge assessment format (predominately multiple-choice questions) in the first 2 years of medical school.

Self-assessment accuracy was quantified in three idiographic (intraindividual) variables developed in our previous work. The bias index is the average difference between each student's estimated performance (x_e) and actual score (x_a) over a series of n observations:

$$\frac{\sum(x_e - x_a)}{n}.$$

This index provides information about the extent to which, on average, students over- or underestimated their performance and by how much.

The deviation index is calculated as the average absolute deviation of the estimated score (x_e) from the actual score (x_a) over n observations:

$$\frac{\sum|x_e - x_a|}{n}.$$

In contrast to the bias index, which allows over- and underestimates to cancel out, the deviation index summarizes how far a student's estimates deviate from actual performance.

The covariation index assesses the correlation between a student's estimated and actual performances over the n observations, i.e. the extent to which variations in a student's estimates parallel variations

in actual performance. Note that the covariation is not influenced by differences between the values of the estimated and actual scores (i.e. bias or deviation scores). We used the Pearson correlation coefficient to quantify covariation.

Statistical methods

Gender and under-represented minority distributions for the three classes were examined using chi square tests. Average medical college admission test (MCAT) scores for the three classes were examined using analysis of variance.

In consideration of the two different operationalizations of stability, the stability of the three self-assessment measures was evaluated in two ways. In the first method, stability of student self-assessment accuracy over the 3-year time frame was examined by a multivariate analysis of variance with repeated measures. These analyses examined the magnitude of self-assessment accuracy over the four intervals.

The second method examined the correlations of each self-assessment accuracy measure between consecutive periods. Only students who had non-missing values for all the periods were used in the analyses. Pearson product–moment correlation was used to evaluate the relationship between period pairs. These analyses examined the extent to which students with relatively high or low levels of self-assessment accuracy in one period were still high or low in the following period.

We treated the data, both analytically and conceptually, in an idiographic (individualized) manner, rather than a more traditional nomothetic (group-based) manner. In other words, each student's data (actual and self-assessed performance) were used to define the self-assessment accuracy of that student. All analyses were done on a 'within-subject' rather than 'between-subject' basis. Group-based outcomes were obtained by averaging individual results. For example, rather than computing 22 correlation coefficients between actual and self-assessed performance on the 22 examinations or tests in the M1 winter term for the 155 students in the 1999 class, we computed 155 individual correlations, one for each student over the 22 examinations. The resulting individual correlations indicated the strength of the covariation between self-assessed performance and actual score for each student, rather than a group-based correlation that does not provide individuating information.

Results

Demographics

Demographic comparisons of the three classes are presented in Table 1. There were no statistically significant differences in the percentage of women, the percentage of minorities and the average MCAT scores among the three classes.

Repeated measures

The multivariate repeated measures analysis of variance indicated that all three self-assessment accuracy meas-

Table 1 Medical student demographics

	Class 1999 ($n = 163$)	Class 2000 ($n = 169$)	Class 2001 ($n = 168$)
Women	42%	38%	40%
Under-represented minority (95% CI)	17% (14·1–19·9)	17% (14·1–9·9)	14% (11·3–16·7)
Medical college admission test score average (95% CI)	10·7 (10·4–11·0)	11·1 (10·8–11·3)	11·1 (10·9–11·3)

Table 2 Means for performance, performance estimates, and self-assessment accuracy measures for each assessment period

Self-assessment measure	*n*	M1 winter term Mean (95% CI)	M2 autumn term Mean (95% CI)	M2 winter term Mean (95% CI)	M3 OSCE Mean (95% CI)
Bias (arithmetic differ.)	343	−2·8 (−3·6 to −2·1)	−2·7 (−3·3 to −2·1)	−2·2 (−2·9 to −1·4)	1·6 (0·7–2·5)
Deviation (absolute differ.)	343	7·8 (7·2–8·3)	7·5 (7·1–8·0)	7·8 (7·3–8·3)	12·9 (12·5–13·4)
Actual-est. covariation	297	0·41 (0·38–0·44)	0·37 (0·34–0·41)	0·36 (0·32–0·40)	0·26 (0·22–0·29)
Self-estimates	343	82·9 (82·2–83·5)	86·2 (85·4–86·9)	85·8 (85·1–86·6)	79·6 (78·9–80·3)
Actual performance scores	388	85·2 (84·6–85·7)	88·3 (87·8–88·8)	87·4 (86·9–87·9)	77·8 (77·2–78·4)

Table 3 Correlations between succeeding assessment periods

		M1 winter & M2 autumn		M2 autumn & M2 winter		M2 winter & M3 OSCE	
Self-assessment measure	*n*	Correlation	(95% CI)	Correlation	(95% CI)	Correlation	(95% CI)
Bias	343	0·63	(0·56–0·69)	0·69	(0·63–0·74)	0·42	(0·33–0·50)
Deviation	343	0·46	(0·37–0·54)	0·55	(0·47–0·62)	0·12	(0·01–0·22)
Covariation	297	0·00	(−0·11–0·11)	0·07	(−0·04–0·18)	0·03	(−0·08–0·14)
Self-estimates	343	0·81	(0·77–0·84)	0·79	(0·75–0·83)	0·36	(0·26–0·45)
Actual performance scores	388	0·60	(0·53–0·66)	0·70	(0·65–0·75)	0·28	(0·19–0·37)

ures, self-estimates and performance scores changed over the course of the study (see Table 2).

The bias scores were negative (indicating an underestimation of actual performance on average) for the first three periods, but became positive in M3 years. This indicated that, on average, the students overestimated their performance on the OSCE. The greatest change for this measure occurred in the M3 years.

Change patterns in the deviation and the covariation values were similar to the bias measure. In the first three periods, scores were relatively consistent, but in the M3 years, the deviation score increased from 7·8 to 12·9 while the mean covariation score decreased from 0·36 to 0·26. The same pattern of change also described the actual self-estimates students provided and the actual performance, both of which showed a decrease in the M3 years.

Correlation between consecutive periods

The correlations between contiguous periods on the three self-assessment accuracy measures indicated that the bias and deviation measures had a similar pattern (Table 3). For both of these measures, the relative stability of students' self-assessment accuracy was moderately high from one period to the next in the first 2 years of medical school, with correlation values ranging from 0·46 to 0·69. However, the correlations between the M2 winter and M3 OSCE periods were substantially lower. The relative stability of the covariation measure was essentially zero between any contiguous periods.

Like the bias and deviation self-assessment measures, the correlations between contiguous periods of actual performance in the first 2 years of medical school are moderately high (0·60 and 0·70), but diminish to 0·28 during the transition to the clinical context. Similarly, the correlations of the self-estimated scores and of the students' actual performance decrease in the same pattern over this period.

Discussion

The means for the performance and self-assessment accuracy measures reflect a fairly high level of stability during the first three assessment periods. However, when self-assessment is required on a different type of task (the third year OSCE), both student performance and self-assessment scores change. For the first time, students overestimated their performance. The increase in both the deviation and covariation scores suggests that self-assessment has become less accurate, which in turn suggests that the type of task or task experience might play a role in making self-assessment judgements.

Although the mean values of the self-assessment accuracy measures changed over time and task, another perspective on the stability of self-assessment accuracy from one time period to the next is reflected in the correlations between assessment periods. The correlations between the M1 and M2 periods vary around 0·65 for the bias measure and 0·50 for the deviation measure (Table 3), accounting for approximately 42% and 25%, respectively, of the variance in scores in the subsequent period. When compared with the stability of actual performance between the same time periods (correlations of approximately 0·65, accounting for approximately 42% of the variance), it is apparent that the stability of self-assessment is similar to the stability of the actual target performance.

The lack of a correlation among the consecutive pairs of the covariation measure likely reflects the relatively low reliability of this measure as calculated for any given time period. Because this correlation of an individual's actual and estimated scores is based on a relatively small number of observations (e.g. 8–22 for a given term and 10–13 in the M3 OSCE), each student's value on this measure is not likely to be very precise or reliable. Thus, the low correlation between consecutive terms may be attenuated by the low reliability of this measure. If so, this may be evidence that this particular measure of self-assessment accuracy is not a very useful

indicator of self-assessment accuracy, in spite of the fact that it quantifies an aspect that is distinct from those summarized in the bias and deviation measures.

The other noteworthy pattern in these results is the decrement in correlation magnitude between the M2 winter and M3 OSCE periods. This decline may be the result of the contrast in tasks reflected in the two periods, such that the classroom-based knowledge assessments of the M2 years do not predict performance or self-assessment accuracy in the clinical performance tasks represented by the OSCE. Note, however, that the same decline in the magnitude of the correlation over this period occurs for actual performance. In fact, the stability of self-assessment accuracy (as indicated by the bias and deviation measures) is at least as great as actual student performance.

Results of this study indicate that medical student self-assessment accuracy is reasonably stable when compared with the stability of actual performance. There may be multiple explanations for the decline in self-assessment accuracy and actual performance between the classroom assessments of knowledge (in the first 2 years) and the clinical assessments of diagnostic and procedural skills (in the OSCE). One is task familiarity. Students who enter medical school have spent years taking paper and pencil examinations. When the task is one in which the students have had limited experience, self-assessment accuracy suffers, as does performance.

An alternative explanation may be that self-assessing one's knowledge (as in the M1 and M2 assessments) is a different process from self-assessing one's performance (as in the OSCE). It may be that self-assessment of knowledge requires dimensions and information that are different from those required in the self-assessment of performance. This judgement process has many dimensions, including the degree an individual understands the task requirements, the accessibility of the targeted competencies to conscious judgement, the evaluation of one's personal skills and resources and past performance on similar tasks. The changing nature of the tasks and the corresponding self-assessment judgements is consistent with the lack of stability in actual performance in these tasks. Performance in the first 2 years predicts only 8% of the variance of performance in the clinical years.

The finding that self-assessment is a reasonably stable characteristic of medical students is a prerequisite for further study of this phenomenon. If self-assessment accuracy had proven to be entirely dependent on task and situation, the search for a conceptual model of self-assessment, pragmatic educational interventions, would become a much more complex endeavour. These results also demonstrate the utility of an idiographic, or intraindividual methodology for studying self-assessment. The focus on individual students and their peculiar strengths and weaknesses constitutes the next stage of research in better understanding the nature and operation of self-assessment in medical education.

Acknowledgements

This research was supported, in part, through a grant from the National Board of Medical Examiner's Stemmler Fund.

References

1 Gordon MJ. A review of the validity and accuracy of self-assessments in health professions training. *Acad Med* 1991;**66**:762–9.
2 Calhoun JG, Woolliscroft JO, Hockman EM, Wolf FM, Davis WK. Evaluating medical student clinical skill performance: Relationships among self, peer, and expert ratings. Proceedings of the 23rd annual Conference on Research in Medical Education. *Res Med Educ* 1984;**23**:205–10.
3 Arnold L, Willoughby TL, Calkins EV. Self-evaluation in undergraduate medical education: a longitudinal perspective. *J Med Educ* 1985;**60**:8.
4 Fitzgerald JT, Gruppen LD, White BA, Davis WK. Medical student self-assessment abilities: accuracy and calibration. *Presented at the Annual Meeting of the American Educational Research Association*, 1997: American Educational Research Association, Chicago, IL.
5 Ward M, Gruppen L, Regehr G. Measuring self-assessment: Current state of the art. *Adv Health Sci Educ* 2002.
6 Fitzgerald JT, Gruppen LD, White C. The stability of student self-assessment accuracy. Presented at the 37th Annual Conference on Research in Medical Education. *Res Med Educ* 1998;21.
7 Gruppen LD, Garcia J, Grum CM. Medical students' self-assessment accuracy in communication skills. *Acad Med* 1997;**72**(**1** Suppl.):S57–9.
8 Gruppen LD, White C, Fitzgerald JT, Grum CM, Woolliscroft JO. Medical students' self-assessments and their allocations of learning time. *Acad Med* 2000;**75**:374–9.
9 Gruppen LD, Baliga S, Fitzgerald JT, White C, Grum CM, Woolliscroft JO, Davis WK. Do personal characteristics and background influence self-assessment accuracy? Presented at the Eighth Ottawa Conference on Medical Education 1998, Philadelphia, PA.
10 Fitzgerald JT, Gruppen LD, White C. The influence of task formats on the accuracy of medical student self-assessment. *Acad Med* 2000;**75**:737–41.

Received 24 *July* 2002; *editorial comments to authors* 3 *October* 2002; *accepted for publication* 10 *December* 2002