# Conditional reliability of admissions interview ratings: extreme ratings are the most informative

R Brent Stansfield[1] & Clarence D Kreiter[2]

CONTEXT Admissions interviews are unreliable and have poor predictive validity, yet are the sole measures of non-cognitive skills used by most medical school admissions departments. The low reliability may be due in part to variation in conditional reliability across the rating scale.

OBJECTIVES To describe an empirically derived estimate of conditional reliability and use it to improve the predictive validity of interview ratings.

METHODS A set of medical school interview ratings was compared to a Monte Carlo simulated set to estimate conditional reliability controlling for range restriction, response scale bias and other artefacts. This estimate was used as a weighting function to improve the predictive validity of a second set of interview ratings for predicting non-cognitive measures (USMLE Step II residuals from Step I scores).

RESULTS Compared with the simulated set, both observed sets showed more reliability at low and high rating levels than at moderate levels. Raw interview scores did not predict USMLE Step II scores after controlling for Step I performance (additional $r^2 = 0.001$, not significant). Weighting interview ratings by estimated conditional reliability improved predictive validity (additional $r^2 = 0.121$, $P < 0.01$).

CONCLUSIONS Conditional reliability is important for understanding the psychometric properties of subjective rating scales. Weighting these measures during the admissions process would improve admissions decisions.

KEYWORDS school admission criteria/*standards; *schools, medical; interviews/*standards; education, medical, undergraduate/ *standards; validation studies [publication type]; observer variation.

*Medical Education 2007; **41**: 32–38*
doi:10.1111/j.1365-2929.2006.02634.x

## INTRODUCTION

Nearly all medical school admissions departments use interviews to assess applicants' non-cognitive skills.[1,2] Admissions departments value this information highly,[3] although it has low reliability and validity.[3–5] Medical school admissions departments rank the impact of interviews on selection decisions above undergraduate grade point average (GPA) and Medical College Admission Test (MCAT) scores, yet only 20% assess predictive validity and only 12% assess reliability.[1] We propose a quantitative method for extracting more reliable and useful information from interview scores for improving their impact on admissions decisions.

Interviewers typically rate applicants on Likert-type scales[1] that yield quantitative, but unreliable, measures.[4,6] Weighted averages of reliable measures (MCAT, undergraduate GPA) with unreliable measures can themselves be unreliable unless weights are chosen carefully.[7] Poorly chosen weights will amplify error and weaken admissions decisions.

There is little evidence for the predictive validity of interviews. The McMaster's admissions interview did not predict licensure examination performance except in the LMCC (Licentiate of the Medical Council of Canada) Part II (Communication) ($r = 0.24$, $P < 0.05$), the same variance predicted by undergraduate GPA and subjective ratings of applicants' essays.[4] Similarly, a private undergraduate

[1]Department of Medical Education, University of Michigan, Ann Arbor, Michigan, USA
[2]Office of Consultation and Research in Medical Education, University of Iowa Carver College of Medicine, University of Iowa, Iowa City, Iowa, USA

*Correspondence*: R Brent Stansfield, Department of Medical Education, University of Michigan, Ann Arbor, Michigan 48109, USA. Tel: 00 1 734 936-1673; Fax: 00 1 734 936-1641; E-mail: rbrent@umich.edu

## Overview

### What is already known on this subject

Likert-like admissions interview ratings are unreliable and have poor predictive validity, but are commonly used because they uniquely measure non-cognitive skills.

### What this study adds

Reliable, valid information exists in low and high ratings, but moderate ratings are uninformative. Quantitative methods for improving the predictive validity of the ratings are examined.

### Suggestions for further research

Further study is required to establish whether 3-point rating scales are more reliable and valid, whether this proposed method would work on rating scales from different interview methodologies, and whether adopting this sort of method during admissions would actually improve cohort performance.

university found admissions interviews predicted undergraduate GPA ($r = 0.10$, $P < 0.05$), but that Scholastic Aptitude Test (SAT) verbal scores explained more variance and interviews explained none of the residual.[5] This may reflect poor measurement of non-cognitive skills by LMCC tests and undergraduate grades, but there are published failures of interviews to predict plainly observable outcomes from non-cognitive factors. For instance, Tel Aviv University began conducting admissions interviews to screen out applicants at risk for dropping out for non-academic reasons; however, the number of such dropouts was unchanged.[8] A New York medical school found the applications of the 35 (of over 1000) students who had developed educationally disruptive psychiatric difficulties were indistinguishable from the rest of students by all measures, including mean interview scores.[9] However, their interview scores contained more inter-rater disagreement. Thus, whereas the interview itself may contain predictive information, average ratings do not.

Unreliable interview scores may not arise from invalid interviewing processes, but rather from the treatment of ratings as homogenously informative measures. Imagine an interviewer able to identify stellar candidates, but unable to distinguish mediocre from poor ones; his high scores would be more informative than his low scores. Despite this validity, his ratings would have low reliability overall. The proper use of his ratings would account for conditional reliability: the reliability of different scale ranges.

We are unaware of any investigations of conditional reliability in admissions interviews. There is only 1 comprehensive review of published reliability assessments of interview ratings.[6] All the cited analyses provide single reliability estimates for the entire scale: ranging from a G-study attributing 9% of interview rating variance to applicant differences,[10] to a Cronbach's alpha of 0.80.[11] Methodological differences may account for the disparities: Cronbach's alpha is a measure of internal consistency, insensitive to the most consequential sources of error in the interview process.[12] Moreover, reliability measures can only estimate the proportion of statistical noise in score variance; they cannot indicate their best use. In their own G-study, Kreiter *et al.*[6] found low-to-moderate reliability with 17–27% of rating variance attributable to applicant differences. They demonstrated that applicants should have multiple interviews, each rated by multiple raters to achieve reliability $\geq 0.6$. An empirical question remains as to whether this low reliability is homogenous across the entire rating scale.

Other investigations of conditional reliability have found heterogeneity of error variance in Likert-type scales. Use of midpoint responses on political opinion questions may represent 'undecided' or 'never thought about it' as opposed to 'neutral' or 'neither agree nor disagree'.[13] This suggests less certainty, and therefore a higher standard error of measurement, in midpoint responses than in non-midpoint responses. A study of education graduate students' responses on an anxiety scale raised Cronbach's alpha from 0.70 to 0.94 merely by treating midpoint responses as missing data.[14] The possibility of varying reliability along the score scale is the reason the American Educational Research Association recommends that 'conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed'.[5]

We used a 2-part analysis to investigate the conditional reliability of interview scores. First, we compared inter-rater disagreement in a real set of interview ratings to those of a randomly generated simulated set. The difference between inter-rater disagreement in the simulated set (which contains no reliability except by chance or mathematical artefact) and the real sets estimates non-artefactual reliability. A non-linear equation is used to model this reliability across the rating scale.

Second, we applied this non-linear equation to a different, independently gathered set of interview ratings of applicants who had matriculated and subsequently completed USMLE (United States Medical Licensing Examination) Steps I and II. We tested the predictive validity on USMLE performance of these interview scores, using the equation as a weighting function.

We tested 2 hypotheses.

1. Inter-rater disagreement relative to chance is not homogenous across the scale. In some regions of the scale, inter-rater disagreement will resemble chance, in others it will be lower.
2. The scale regions with less inter-rater disagreement will show more predictive validity. The correlation of these scores with an estimate of communication skills in clinical settings will be stronger than that of the less reliable regions of the scale.

## METHODS

### Participants: observed and simulated

*Observed set 1*

A total of 478 applicants were selected for interview at the University of Iowa's medical school in 2003 by a minimum cut score on the MCAT and GPA. Two raters were randomly assigned from a large pool of over 115 volunteer faculty members to rate each applicant. A total of 37 applicants were excluded for missing data (at least 1 item rating from at least 1 rater), yielding a final sample size of 441 applicants.

*Observed set 2*

Observed set 2 included 91 applicants who matriculated and completed USMLE Steps I and II at the same institution in 2001, and who had participated in an identical interview.

Applicants from observed sets 1 and 2 participated in identical interview processes. Each applicant responded to 6 items: 5 structured items (the same for all) and 1 unstructured item (of the raters' choosing). Raters independently rated the response to each item and then rated the applicant overall using a 5-point Likert scale, where points 1, 3 and 5 were anchored with a short description of what the responses should contain in reference to each question. Each applicant received 14 ratings: 7 from each of 2 raters. The few non-integer (i.e. 3.5) ratings were rounded up. For each applicant, 1 rater was randomly designated 'rater A' and the other 'rater B'.

*Simulated set*

The simulation was generated using a Monte Carlo method by an SAS script (Version 8.2). We calculated the response frequencies for each scale level for each of the 7 items from all raters and all applicants in observed set 1. These frequencies were used as cut-off scores to translate random numbers into 5-point scale ratings: for instance, of the observed ratings on the unstructured item, 0.2% were rated as 1, 3.1% as 2, 24.9% as 3, 43.9% as 4, and 27.9% as 5, so a random number (ranging from 0 to 1) $\leq 0.002$ became a 1, a random number $< 0.033$ became a 2, a number $< 0.282$ became a 3, and so on. We simulated 100 000 applicants being rated, each by 2 raters. Thus, the overall response distribution (means, standard deviations, skewness) for each item of the simulated sample was identical to that of observed set 1.

### Analysis

The simulated set served as a control for reliability due to chance and mathematical artefact. As this set was modelled on the frequency distribution of ratings on the items of observed set 1, a direct comparison of these 2 datasets' reliability and conditional reliability provides an estimate of the non-artefactual reliability. As the simulated set was not constructed using any knowledge of observed set 2, any gains in reliability or predictive validity of observed set 2 scores from any model generated from observed set 1 and the simulated set will be independently valid of the estimation methods used.

To test for heterogeneity of rating scale reliability, we used inter-rater disagreement. We modelled inter-rater disagreement across the rating scale for observed sets 1 and 2 from the simulated set using non-linear regression by rater A's ratings. A model of heterogeneity of reliability was constructed as the difference between the inter-rater disagreement regression curves for the simulated set (chance agreement) and observed set 1 (observed agreement). This difference served as a weighting function to predict USMLE scores by the mean rating of raters A and B in observed set 2. We predicted USMLE Step I and Step II scores, and also the residual of a regression of Step II by Step I. This residual represents variance in the more non-cognitive skill-based Step II controlling for students' performance on the strictly knowledge-based Step I.

## RESULTS

### Observed set 1 is more reliable than the simulated set

Ratings in observed set 1 were negatively skewed. Every item had a strong mode of 4 and few ratings of 1 or 2. Item means (across all 7 items) were computed for each applicant, for each rater in both observed set 1 and the simulated set. These item means were used for all subsequent analyses.

Although the descriptive statistics of the individual items in the simulated set were identical to those of observed set 1, the distributions of across-item means within student and rater were different, suggesting more reliability in observed set 1: the means of raters A and B in observed set 1 (mean = 3.912) and the simulated set (mean = 3.921) were practically the same, but the standard deviation (SD) of the simulated set (SD = 0.3) was half that of observed set 1 (SD = 0.6). This suggests a stronger effect of central tendency in the simulated set; real raters are more likely to show interitem consistency. This is further supported by the less skewed (skewness = – 0.06) and much more leptokurtic (kurtosis = 2.43) distribution in the simulated set (observed set 1 skewness = – 0.25, kurtosis = – 0.31). Real raters had high interitem reliability (Cronbach's alpha = 0.862), while simulated raters had none (alpha = 0.004).

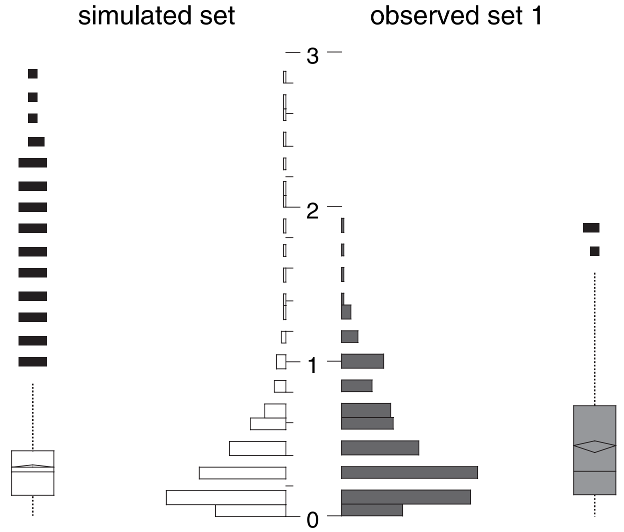

**Figure 1** Comparative histogram of absolute inter-rater disagreement in the simulated set (on left, in white) and observed set 1 (on right, in grey).

Despite this, interrater disagreement (the absolute value of rater A's minus rater B's item means) did not differ dramatically between the 2 sets (Fig. 1). The means and SDs of the disagreements were comparable (observed set 1: mean = 0.45, SD = 0.34; simulated set: mean = 0.33, SD = 0.30).

Although the mean disagreement of the observed raters was higher than that of the simulated set, simulated raters disagree more. Below the median, inter-rater disagreements were identical, but above the median, observed set 1 disagreements are much lower. The observed set's lack of large disagreements indicates reliability: observed raters tend to disagree less than they might by chance alone.

### Low and high ratings are more reliable

Figure 2 plots inter-rater disagreement by rater A's item mean for all 3 datasets. In each, when rater A gives low or high ratings, rater B disagrees more than when rater A gives moderate ratings. This is due in part to range restriction: higher inter-rater differences are possible when 1 rater uses an extreme value. This relationship in the simulated set estimates how much disagreement is expected by range restriction alone. A regression model quantifies this estimate. A linear fit is weak ($r^2 = 0.001$); a quadratic fit is much more successful ($r^2 = 0.312$). Additional parameters do not improve the fit: a cubic fit yields

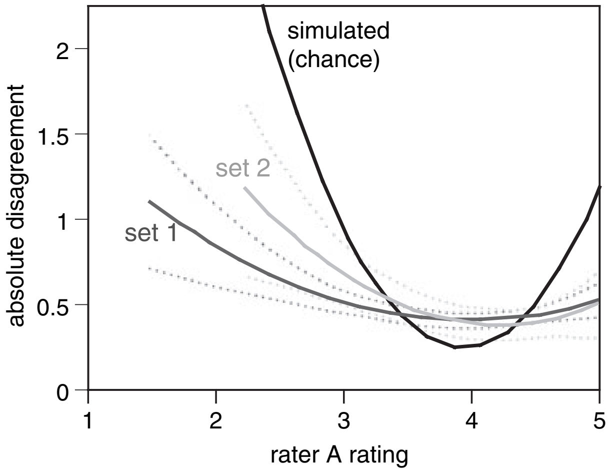

**Figure 2** Quadratic relationship of absolute value of inter-rater disagreement and rater A rating for the simulated set (black), and observed sets 1 (dark grey) and 2 (light grey). Dotted curves show the 95% confidence interval.

$r^2 = 0.314$; a quartic fit yields $r^2 = 0.327$. The quadratic fit is adopted for this analysis.

The relationship in both observed sets is similarly quadratic, but flatter than that of the simulated set, showing less interrater disagreement at the scale ends. When rater A scores an applicant higher than about 4.5 or lower than about 3.2, rater B tends to agree with A more than is expected by chance alone. When rater A scores an applicant at between 3.6 and 4.2, rater B tends to agree with A less than expected. This is not an artefact of the regression model: mean inter-rater disagreement at all discrete levels of rater A's item means are higher in both observed sets than in the simulated set.

The difference between these curves represents the amount of agreement not due to chance alone, an estimate of conditional reliability. Subtracting the quadratic fit of observed set 1 from that of the simulated set yields a quadratic function of item mean score *s* (ranging from 1 to 5); we label the result of this function *d*. Negative values of *d* represent the middle region where the observed interrater disagreement is greater than what we might expect by chance; we chose to treat these values as zero:

$$d = \text{maximum of } (10.594 - 5.488s + 0.700s^2)\text{or } 0 \quad (1)$$

Restricting *d* to positive values makes it estimate the likelihood a different rater will actually (non-artefactually) agree with a given score. We tested whether this value magnifies any predictive validity in the interview ratings.

### Weighting low and high responses improves validity

In observed set 2, mean ratings predict neither USMLE Step I scores ($r = -0.149$, not significant), nor Step II ($r = -0.069$, not significant). A stepwise regression of Step II scores found that mean ratings did not predict Step II variance not explained by Step I scores: entering Step I scores yielded $r^2 = 0.475$, entering mean ratings raised it to $r^2 = 0.476$ (a non-significant increase).

We computed *d* using Equation 1 using the average of raters A's and B's item means as *s*, then used *d* as weights in the same correlation and regression models. In these weighted analyses, the impact of each datapoint is a function of its *d* score. Ratings with $d = 0$ are ignored, and the rest have impact proportional to their *d* scores. *d*-weighted mean ratings did not positively predict scores on Step I ($r = -0.255$, not significant) or Step II ($r = 0.163$, not significant). The *d*-weighted stepwise regression of Step II scores again found a strong effect of Step I scores ($r^2 = 0.463$), and entering mean ratings explained significantly more variance ($r^2 = 0.584$, a significant rise; $P < 0.01$).

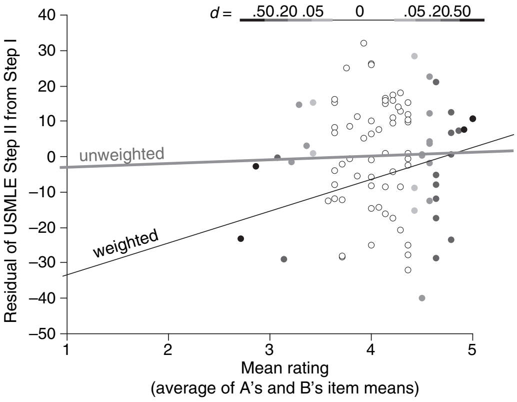

**Figure 3** US Medical Licensing Examination Step II residuals from Step I scores in observed set 2, explained by mean rating using an unweighted (grey line) or weighted (black line) regression model. Empty circles are applicants with $d = 0$; filled circles are applicants with $d > 0$; darker circles indicate higher *d*-values.

Figure 3 shows the regression results using the unweighted (grey line) and *d*-weighted (black line) models. Only 34 of the 91 applicants had *d*-values > 0, so the weighted model has a much smaller sample size, but shows a much stronger trend.

## DISCUSSION

Raters tend to agree more about the lowest and highest quality applicant interviews. This agreement is not a mathematical artefact: the simulated set contains much more inter-rater disagreement at extreme ratings than observed sets 1 or 2 (Fig. 2). Raters tend to disagree more than chance about applicants whom 1 rater has deemed average. These moderate ratings are actually 'negatively reliable', suggesting an invalid use of the modal response, perhaps denoting 'I don't know' rather than 'average applicant'. If so, these large inter-rater disagreements reflect differences in confidence rather than substance. As raters rarely use levels 1 and 2, the modal level 4 is effectively the midpoint on a 3-point scale; these results mirror those finding midpoint responses unreliable.[13,14] Our *d*-weighted regression analysis effectively treats moderate ratings as missing data improving validity.

The fact that the *d*-weighted regression was more successful than the unweighted one is powerful given that the equation for *d* was derived from a different dataset. A second simulated dataset based on observed set 2 might yield a different equation for *d*, with even more predictive validity. A generalised equation for *d*, incorporating many independent sets of interview data, would likely yield highly stable estimates of conditional reliability.

The above methods are computationally expensive, but there are easy ways of approximating them. A 3-point scale ('unacceptable', 'outstanding' or 'neither/unsure') might be at least as informative as the 5-point scale analysed here, if raters used it similarly to the way these raters used 3, 4 and 5. One could correlate any existing Likert-like ratings with students' USMLE Step II scores (for instance), treating modal responses as missing or not. The quantitative method yielding the best correlation might also yield better admissions decisions.

More importantly, these results suggest that ignoring moderate interview ratings entirely during the admissions process is preferable to using them when computing larger weighted sum scores. Introducing unreliable measures into weighted averages with reliable ones can compromise the reliability of the resulting score.[6] Treating all moderate responses as missing data eliminates the impact of the noise in those responses, while allowing extreme scores (which in these data have some predictive validity) to influence applicants' relative standings.

The equation for *d* in this analysis generalised from observed set 1 to observed set 2 probably because the interview processes were identical. Equation 1 is unlikely to yield the same results for another interview methodology at another institution. It provides a continuous quantification of the certainty of interview scores, allowing those scores a continuous degree of importance, which is more informative than making a simple missing/non-missing distinction. Transforming the rating scale by a *d*-like equation (making low ratings more negative, high ratings more positive, and moderate ratings more zero-like) before computing a weighted average of all applicant information would be easy. In this way moderate scores would have no impact on the weighted average ($d = 0$), whereas non-moderate scores would impact the overall score to the degree of their extremity.

Using only the informative ranges of any measure will reduce the amount of statistical noise, improve reliability and possibly validity, and allow for better admissions decisions. Any investigation of the reliability of measures used for such important decisions as medical school admissions is admirable, but consideration of conditional reliability can improve the effectiveness of those decisions even more.

*Contributors:* both authors contributed to the theoretical concept of this study and to the interpretation of data. RBS carried out the analysis design, data analysis and the literature review and wrote the paper. CDK gathered the data and contributed to revisions of the manuscript.

*Acknowledgements:* the authors thank Catherine Solow of the University of Iowa Medical School Admissions Department.
*Funding:* none.

*Conflicts of interest:* none.
*Ethical approval:* this study was approved by the University of Iowa Institutional Review Board.

## REFERENCES

1 Johnson EK, Edwards JC. Current practices in admission interviews at US medical schools. *Acad Med* 1991;**66**:408–12.
2 Puryear JB, Lewis LA. Description of the interview process in selecting students for admission to US medical schools. *J Med Educ* 1981;**56**:881–5.
3 Kulatunga-Moruzi C, Norman GR. Validity of admissions measures in predicting performance outcomes: the contribution of cognitive and non-cognitive dimensions. *Teach Learn Med* 2002;**14**:34–42.
4 Mitchell K, Haynes R, Koenig J. Assessing the validity of the updated Medical College Admission Test. *Acad Med* 1994;**69**:394–401.
5 Shahani C, Dipboye RL, Gehrlein TM. The incremental contribution of an interview to college admissions. *Educ Psychol Meas* 1991;**51**:1049–61.
6 Kreiter CD, Yin P, Solow C, Brennan RL. Investigating the reliability of the medical school admissions interview. *Adv Health Sci Educ Theory Pract* 2004;**9**:147–59.
7 Kreiter CD, Gordon JA, Elliott S, Callaway M. Recommendations for assigning weights to component tests to derive an overall course grade. *Teach Learn Med* 2004;**16**:133–8.
8 Ehrenfeld M, Tabak N. Value of admission interviews in selecting of undergraduate nursing students. *J Nurs Manag* 2000;**8**:101–6.
9 Keill SL, Willer B. Detection of psychiatrically at-risk applicants in the medical school admission process. *J Med Educ* 1985;**60**:800–2.
10 Harasym PH, Woloschuk W, Mandin H, Brundin-Mather R. Reliability and validity of interviewers' judgements of medical school candidates. *Acad Med* 1996;**71** (Suppl):40–2.

11 Tutton PJ. Medical school entrants: semistructured interview ratings, prior scholastic achievement and personality profiles. *Med Educ* 1993;**27**:328–36.

12 Schmitt N. Uses and abuses of coefficient alpha. *Psychol Assess* 1996;**8**:350–3.

13 Raaijmakers QAW, van Hoof A, 't Hart H, Verbogt TFMA, Vollebergh WAM. Adolescents' midpoint responses on Likert-type scale items: neutral or missing values? *Int J Public Opin Res* 2000;**12**:208–16.

14 Weems GH, Onwuegbuzie AJ. The impact of midpoint responses and reverse coding on survey data. *Meas Eval Counsel Dev* 2002;**34**:166–76.

15 American Educational Research Association. *Standards for Educational and Psychological Testing, Standard 2.14.* Washington, DC: AERA 1999;35.

*Received 4 May 2006; accepted for publication 17 July 2006*