Evaluating a Structured and Unstructured Format for the Medical School Preadmission Interview

Introduction

The medical education researcher literature consistently recommends a structured format for the medical school preadmission interview (MSPI). There is however little direct evidence for making this recommendation.

A recent study by Kreiter et al. (2006) suggests that there is not a reliability-related rationale for presenting the same questions to all examinees.

However, using identical questions is only one facet of interview structure and there are a wide range of interview characteristics that collectively determine the degree of MSPI structure. In fact, researchers investigating both employment and admission types of interviews have developed a continuum classification scheme for level of interview structure. Unfortunately, there is little research evidence regarding how reliability is impacted by the level of MSPI interview structure. This study examines the contribution of a structured and unstructured MSPI conducted at the University of Iowa.

Method

Interview Description

As part of the CCOM application process, particularly well-qualified candidates are selected by the Admissions Committee to participate in a 25 minute interview with two faculty members. There are two portions to the interview: 1) structured format – where candidates are read predetermined questions and their responses are scored on a scale from 1-5 (with 5=excellent) using an established scoring rubric; and 2) unstructured format - where there is a free-flowing exchange between faculty and the candidates about any topics of interest to the faculty and/or candidates. Scores ranging from 1-5 are also awarded on the unstructured portion of the interview, but are not guided by specific scoring rules. An overall score ranging from 1-5 is also awarded based on the interviewer's overall impression of the candidate.

Research Questions

1. Are the scores on the structured and unstructured interview components equally reliable? (inter-rater and test-retest)
2. Do the structured and unstructured interview components measure the same constructs?
3. Are there technical advantages to combining the interview components into an aggregate measure?

Data

CCOM application data from 2003-2007 was compiled for those candidates who were interviewed one or more times. The counts of applicants by admission outcomes for each application year are shown below.

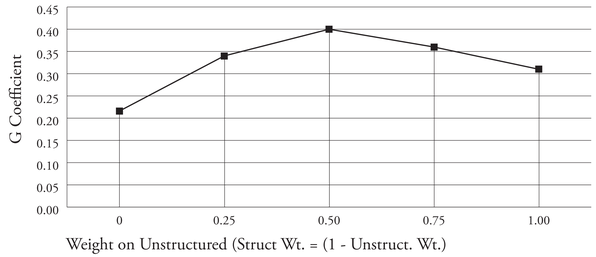
Application Year * Admission_grp Crosstabulation

		Admission_grp				
		Admitted/ enrolled	Admitted/ not enrolled	Rejected	Withdrew before eval completed	Total
Year	2003	142	134	203	68	547
	2004	142	150	225	71	588
	2005	142	156	269	78	645
	2006	142	132	259	77	610
	2007	148	155	295	83	681
Total		716	727	1251	377	3071

The average scores for each interview component are displayed by admission outcome in the following table.

Admission_grp		Mean unstructured	Mean overall	Mean Structured
Admitted and enrolled	Mean	4.3469	4.2050	4.0289
	N	708	711	711
	Std. Deviation	.55734	.58750	.46667
Admitted but did not enroll	Mean	4.4147	4.2579	4.0972
	N	727	727	727
	Std. Deviation	.54430	.57543	.45793
Rejected	Mean	3.6908	3.5332	3.5946
	N	1251	1251	1251
	Std. Deviation	.79742	.74483	.56143
Withdrew before eval completed	Mean	4.0703	3.9105	3.8407
	N	376	377	377
	Std. Deviation	.61161	.60947	.49364
Total	Mean	4.0610	3.9072	3.8447
	N	3062	3066	3066
	Std. Deviation	.74308	.73282	.55416

Composite Reliability of One Interview as a Function of Unstructured Weight



Description of the analysis

Four generalizability (G) studies were conducted on the data described. The first two studies, using the data from all 3043 interviews, were designed to estimate and compare levels of inter-rater reliability, or how well raters agreed within the structured and unstructured component of the interview. Both of these G studies used a simple rater-nested-within-person [r:p] random model. The first implementation of this model used just the data from the structured part of the interview and the second r:p analysis used only the data from the unstructured ratings. The third G study used the rating data from both the structured and unstructured parts of the interview for all 3034 applicants in a rater-nested-within-person-crossed-with-format [r:p x f] mixed model design. Format in this model is fixed and has two levels (structured and unstructured) allowing an estimate of the person-by-format interaction (pf). The magnitude of the pf variance component (VC) provides information on the degree to which the two formats assess different aspects of examinee performance. The fourth G study used data from those applicants interviewed twice (n = 163) and employed a multivariate design with a person-by-interview model for applicants interviewed in two separate years with both person and interview having observations for each level of the fixed format facet.

3 univariate G studies and 1 multi-variate G study

1. Inter-rater reliability for structured [r:p]

G Study Results

Effect	DF	VC	SE
P	3042	0.2033609	0.0082646
R : P	3043	0.2047667	0.0052479

Alpha = .67

2. Inter-rater reliability for unstructured [r:p]

G Study Results

Effect	DF	VC	SE
P	3042	0.4049061	0.0146039
R : P	3043	0.2917105	0.0074761

Alpha = .74

3. Variance components for a single interview [r:p x f]

r:p x t

Effect	DF	VC	SE	%
P	3042	0.2708625	0.0095719	40
R : P	3043	0.1824992	0.0046772	27
F	1	0.0200223		3
PF	3042	0.0665419	0.0037862	9
RF : P	3043	0.1314788	0.0033696	20

4. Variance components and reliability obtained from two complete interviews each employing a structured and unstructured format. [p•x i•]

Estimated G Study Variance and Covariance Components

Effect	Structured	Unstructured
P	0.06479	0.82227
	0.09234	0.19466
O	0.00104	
	0.00447	0.01077
PO	0.23586	
	0.21466	0.43996

Conclusions

The unstructured format was more reliable from both an inter-rater rater agreement perspective and in the random replications (test/retest) analysis. In addition, the universe score correlation (ru = .82) and person-by-format interaction indicated that the two formats do not measure identical aspects of applicant performance. An examination of weighted composite scores indicates a sum score with approximately equal weights on both formats maximizes reliability and information.

Authors: R Axelson, J McGovern, C Kreiter, K Ferguson, C Solow, and K Huebner