Comparative Reliability of Structured Versus Unstructured Interviews in the Admission Process of a Residency Program

Danielle Blouin, MD, MHPE
Andrew G. Day, MSc
Andrey Pavlov, PhD

Abstract

Background Although never directly compared, structured interviews are reported as being more reliable than unstructured interviews. This study compared the reliability of both types of interview when applied to a common pool of applicants for positions in an emergency medicine residency program.

Methods In 2008, one structured interview was added to the two unstructured interviews traditionally used in our resident selection process. A formal job analysis using the critical incident technique guided the development of the structured interview tool. This tool consisted of 7 scenarios assessing 4 of the domains deemed essential for success as a resident in this program. The traditional interview tool assessed 5 general criteria. In addition to these criteria, the unstructured panel members were asked to rate each candidate on the same 4 essential domains rated by the structured panel members. All 3 panels interviewed all candidates. Main outcomes were the overall, interitem, and interrater reliabilities, the correlations between interview panels, and the dimensionality of each interview tool.

Results Thirty candidates were interviewed. The overall reliability reached 0.43 for the structured interview, and 0.81 and 0.71 for the unstructured interviews. Analyses of the variance components showed a high interrater, low interitem reliability for the structured interview, and a high interrater, high interitem reliability for the unstructured interviews. The summary measures from the 2 unstructured interviews were significantly correlated, but neither was correlated with the structured interview. Only the structured interview was multidimensional.

Conclusions A structured interview did not yield a higher overall reliability than both unstructured interviews. The lower reliability is explained by a lower interitem reliability, which in turn is due to the multidimensionality of the interview tool. Both unstructured panels consistently rated a single dimension, even when prompted to assess the 4 specific domains established as essential to succeed in this residency program.

Background

Interviews have long been regarded as one of the most important screening tools in resident selection.[1–11] Although the reliability of structured interviews (SIs) is consistently reported as higher than that of unstructured interviews (UIs), we could find no study directly comparing both types of interviews in a common pool of interviewees.[4,12–17]

Studies of the predictive validity of interviews for resident clinical performance have found mixed results.[18–22] The interviews in these studies were not structured, and only one study[21] reports an interrater reliability (another study[18] describes the interviews as structured, but the description does not support this statement). Although the predictive validity of interviews is the ultimate goal, their reliability has to be optimized first before they can be used in a predictive manner.

In 2006, we designed a highly structured interview for applicants to a residency program in emergency medicine (EM).[23] The tool yielded a good, not excellent, reliability (generalizability coefficient, 0.67).[23] Given the efforts involved in structuring the interview, we sought to confirm

All authors are at Kingston General Hospital in Kingston, Ontario, Canada. **Danielle Blouin, MD, MHPE,** is Associate Professor in the Department of Emergency Medicine, Queen's University; **Andrew G. Day, MSc,** is Senior Biostatistician at the Clinical Research Centre; and **Andrey Pavlov, PhD,** is Assistant Biostatistician at the Clinical Research Centre.

Funding: This study was made possible through a grant from the Clinical Teachers Association at Queen's University Endowment Fund.

Corresponding author: Danielle Blouin, MD, MHPE, Department of Emergency Medicine, Kingston General Hospital, 76 Stuart Street, Kingston, ON K7L 2V7 Canada, 613.548.2368, blouind@kgh.kari.net

Received December 20, 2010; revision received June 20, 2011; accepted June 25, 2011.

DOI: http://dx.doi.org/10.4300/JGME-D-10-00248.1

that its reliability was indeed higher than that of the UIs traditionally used by our program. Our study compares the reliability of SIs and UIs when applied to a common pool of applicants to an EM residency program.

What was known

Structured interviews are reported to have higher reliability in evaluating applicants. No direct comparisons have been performed.

What is new

Structured interviews have lower reliability due to their multidimensional nature. This may add validity and discriminant ability.

Limitations

Small sample, single site, and single specialty.

Bottom line

Structured interviews offer a more valid assessment by discriminating between different dimensions, but require a greater number of interviewees and scenarios for comparable reliability.

Methods

Study Design

In this prospective study, an SI was added to 2 UIs traditionally used in the admission process of a residency program. Candidates were interviewed by all 3 panels. Each panel comprised 3 interviewers (2 faculty members and 1 senior resident). The Research Ethics Board at Queen's University approved this study (REB No. EMED-083-06).

Study Setting and Population

The study took place at Queen's University, Kingston, ON, Canada. The study focused on the 2008 applicants to the 5-year EM program accredited by the Royal College of Physicians and Surgeons of Canada.

Study Protocol

The SI development has been previously reported.[23] In summary, the tool (available upon request) consisted of 7 clinical scenarios exploring 4 dimensions of performance deemed essential for EM residents: professionalism (2 scenarios), teamwork (2 scenarios), maturity (2 scenarios), and patient advocacy (1 scenario). These dimensions were extracted using the critical incident technique.[24,25] Each scenario asked for a decision or an approach, and its rationale, rated on a 5-point scale with anchors describing examples of worst, neutral, and best responses. The tool ended with a global assessment of the candidate's suitability for our EM program using a 10-point Likert scale with 3 anchors: "Worst candidate ever"; "Average"; and "Best candidate ever." Rater training consisted of an individual and collective review of the papers of Campion et al[16,17] on SI features, and ratings of videotaped mock candidate interviews followed by discussion until adequate standardization was achieved.

The unstructured panels used the program traditional rating form and marked candidates on 5 criteria: (1) general presentation, (2) character (honesty/confidence/energy), (3) quality of answers (organization/thoughtfulness), (4) suitability for EM specialty, and (5) personality (suitability to our EM program). Each criterion was scored on a scale of 1 to 10, with 1 "very weak," 5 "average," and 10 "very strong." For study purposes, the UI members also rated candidates on the 4 domains marked by the structured panels (professionalism, etc) using a 5-point scale without specific descriptive anchors.

Candidates' files contained the standard written documentation requested by the Canadian Residency Matching Service.[26] Approximately 2 months before the interviews, UI members individually reviewed and scored applicants' files, then met to create the list of candidates to interview. As per the recommendations of Campion et al,[16] SI members did not preview files and did not participate in the selection process. All 3 panels remained constant for all candidates. Each interview lasted 20 minutes.

The SI proposed the same scenarios to each candidate by the same interviewer. Interviewers alternated in presenting scenarios throughout the interview. Each interviewer scored the candidates on every scenario. Interviewers did not prompt further response unless specifically required to by the scoring form, nor did they ask exploratory questions. Candidates could pose questions only after the scoring was completed. Interviewers were encouraged to take notes and were specifically asked not to discuss candidates, answers, or assigned scores between interviews.

The UI members did not receive question scripts, conducted conversational interviews, and rated each candidate on the 5 traditional criteria and on the 4 domains. To avoid interfering with the traditional interview process, UI members were not specifically precluded from discussing candidates between interviews; however, they had no time to share their ratings because of the tight scheduling.

Key Outcomes Measures

Main outcomes were the interitem, interrater, and overall reliability, the correlation between panels, and the dimensionality of each interview panel. The correlation between panels was measured using summary scores (sum of all scores candidates received from each rater, across items).

TABLE 1 **Components Variance Estimates: Unstructured Panels Using Traditional Form**[a]

	Unstructured Panel 1		Unstructured Panel 2		Structured Panel	
	Estimated Variance σ^2	SE σ^2	Estimated Variance σ^2	SE σ^2	Estimated Variance σ^2	SE σ^2
Subject (s)	79.0	25.9	55.5	21.0	45.1	30.0
Rater (r)	17.4	19.2	13.5	16.0	2.3	4.7
Item (i)	4.1	9.2	3.1	3.1	52.9	37.9
Subject*rater (sr)	40.1	3.7	56.5	11.8	30.9	10.5
Subject*item (si)	11.3	3.8	9.6	3.0	281.2	36.5
Rater*item (ri)	0.7	1.1	1.7	1.5	3.5	3.8
Error (sir)	46.3	4.3	35.1	3.3	171.6	13.0
Interitem reliability (IIR)	0.97	0.03	0.97	0.03	0.53	0.16
Interrater reliability (IRR)	0.83	0.09	0.73	0.11	0.82	0.09
Generalizability coefficient ($E\rho^2$)	0.81	0.09	0.71	0.10	0.43	0.16

[a] $IIR = \sigma^2(s) / [\sigma^2(s) + \sigma^2(si) / n_i]$
$IRR = [\sigma^2(s) + \sigma^2(si) / n_i] / [\sigma^2(s) + \sigma^2(si) / n_i + \sigma^2(sr) / n_r + \sigma^2(sir) / n_i n_r]$
$E\rho^2 = IIR * IRR = \sigma^2(s) / [\sigma^2(s) + \sigma^2(si) / n_i + \sigma^2(sr) / n_r + \sigma^2(sir) / n_i n_r]$
where $n_r = 3$ raters and $n_i = 5$ and 7 items for unstructured and structured panels, respectively.

TABLE 2 **Components Variance Estimates: Unstructured Panels Rating the Same 4 Domains as the Structured Panel**[a]

	Unstructured Panel 1		Unstructured Panel 2	
	Estimated Variance σ^2	SE σ^2	Estimated Variance σ^2	SE σ^2
Subject (s)	80.2	45.8	23.4	14.5
Rater (r)	25.5	37.0	35.4	38.8
Item (i)	0.5	4.7	0.0	NA
Subject*rater (sr)	214.4	45.9	64.5	15.5
Subject*item (si)	6.7	5.9	1.5	4.7
Rater*item (ri)	9.7	7.3	2.2	2.2
Error (sir)	76.9	8.7	71.8	7.8
Interitem reliability (IIR)	0.98	0.08	0.98	0.12
Interrater reliability (IRR)	0.51	0.19	0.46	0.21
Generalizability coefficient ($E\rho^2$)	0.50	0.18	0.46	0.20

[a] $IIR = \sigma^2(s) / [\sigma^2(s) + \sigma^2(si) / n_i]$
$IRR = [\sigma^2(s) + \sigma^2(si) / n_i] / [\sigma^2(s) + \sigma^2(si) / n_i + \sigma^2(sr) / n_r + \sigma^2(sir) / n_i n_r]$
$E\rho^2 = IIR * IRR = \sigma^2(s) / [\sigma^2(s) + \sigma^2(si) / n_i + \sigma^2(sr) / n_r + \sigma^2(sir) / n_i n_r]$
where $n_r = 3$ raters and $n_i = 4$ items.

TABLE 3 **Interitem Consistency for Unstructured Interview Panels**

	Cronbach α[a]
Unstructured panel 1: rater 1	0.90
Unstructured panel 1: rater 2	0.96
Unstructured panel 1: rater 3	0.86
Unstructured panel 2: rater 1	0.90
Unstructured panel 2: rater 2	0.94
Unstructured panel 2: rater 3	0.93
Structured panel: rater 1	0.48
Structured panel: rater 2	0.61
Structured panel: rater 3	0.53

[a] On 5 items for unstructured panels A and B; on 7 items for structured panel.

For the UIs, the reliability analysis was performed on the marking of the 5 criteria traditionally rated as well as the 4 essential domains.

Data Analysis

Pearson correlation coefficient was used to measure the correlation between the summary score of each panel.

Generalizability theory[27] was used to compute the reliability. This approach decomposes the total variance of the overall assessment into components due to subjects, raters, items, and all interactions between these terms. These components are then used to construct reliability measures. We present the interitem, the interrater, and the overall reliability (generalizability) (see TABLES 1 and 2 footnotes for formulae).[27] The variance components (VCs) and their standard errors were estimated by restricted maximum likelihood implemented in the MIXED procedure of SAS 9.2 (SAS Institute Inc, Cary, NC), assuming that subject, rater, and items were each random effects. Standard errors for the reliability coefficient were estimated by the standard deviation of 1000 bootstrap samples, where each bootstrap sample consisted of a random resampling of items, raters, and subjects. For the UI analyses, ''items'' were the 5 criteria traditionally rated by the interviewers; in the second analysis, ''items'' represented the additional 4 domains rated by the interviewers. For the SI, the 7 scenarios constituted the ''items.'' All ratings were scaled between 0 (worst possible score) and 100 (best possible score). This standardized scaling has no impact on the reliability coefficients, but facilitates comparison of VCs between panels. Finally, the 3 reliability coefficients presented in this study are each based on relative error variance rather than absolute error variance because the focus of this manuscript is the relative ranking of candidates rather than their actual scores.[27] Interitem consistency was measured using Cronbach alpha.

Factorial analysis using the principal component method and a covariance matrix was performed to extract factors with an Eigen value of at least 1. Varimax rotation with Kaiser normalization method was chosen.

Results

In February 2008, our program interviewed 30 candidates during 2 days. The summary measures from the 2 UIs were significantly correlated (Pearson correlation coefficient, $r = 0.53$ [$P = .003$]), but neither UI correlated with the SI ($r = 0.14$ [$P = .46$] and $r = -0.25$ [$P = .19$]).

The SI interrater reliability was 0.82, compared with 0.83 and 0.73 from the UIs using the traditional items; the corresponding interitem reliabilities were 0.53, 0.97, and 0.97. Their product, overall reliability (generalizability coefficient), was 0.43 for the SI, and 0.81 and 0.71 for the UIs using traditional items (TABLE 1).

The UI overall reliability coefficients when rating candidates on the 4 domains were 0.50 and 0.46, with the 2 panels having interitem reliabilities of 0.98 but interrater reliabilities of 0.51 and 0.46, respectively (TABLE 2).

Rater-specific interitem consistency, measured by Cronbach α, varied from 0.86 to 0.96 for the UIs, and from 0.48 to 0.61 for the SI (TABLE 3). The SI assessed between 2 and 3 dimensions; UIs were unidimensional (TABLE 4).

Discussion

Overall reliability is a product of interitem reliability (how closely candidates' scores match across scenario/criterion) and interrater reliability (how closely candidates' scores match across interviewers for a particular scenario/criterion). A good interview tool demands a high interrater reliability. A high interitem reliability suggests that all items assess a single domain (either a single domain is being assessed or raters cannot discriminate between proposed domains).

Our results fail to confirm a higher overall reliability of SIs over UIs when applied to the same pool of candidates. Both UIs achieved a moderate to high reliability, whereas that of the SI is quite poor.

The SI low reliability is explained by a low interitem reliability (VC, 281.2); that is, candidates scored differently across scenarios (TABLE 1). The interrater variance is small (VC, 30.9), implying that the relative scoring of candidates was consistent between raters. The reverse situation exists for both UIs, with the variance between raters being the main component, implying poor agreement between raters but consistent scores across criteria (ie, poor discrimination between the criteria purportedly assessed).

TABLE 4 **Number of Factors With an Eigen Value Greater Than One for Each Interviewer**

	No. of Factors	
	From Traditional 5 Criteria Tool	Additional 4 Domains Tool
Unstructured panel 1: rater 1	1	1
Unstructured panel 1: rater 2	1	1
Unstructured panel 1: rater 3	1	1
Unstructured panel 2: rater 1	1	1
Unstructured panel 2: rater 2	1	1
Unstructured panel 2: rater 3	1	1
Structured panel: rater 1	3	
Structured panel: rater 2	3	
Structured panel: rater 3	2	

This is a critical finding because the ultimate purpose of the SI is to decrease the rater variability. Our SI accomplished this.

The poor discrimination of the UIs is substantiated by their high Cronbach α, suggesting that the 5 criteria assessed actually measure a single dimension (TABLE 3). In contrast, the SI has a low Cronbach α and appears to be multidimensional. A formal factor analysis was conducted to determine how many domains were assessed by each interview type (TABLE 4). One factor is extracted for both UIs, with 2 and 3 factors for the SI. The SI multidimensionality explains the high interitem variance. The UIs rated candidates on only 1 dimension, despite being asked to rate on 5 criteria. The 1 dimension rated appears to be the same across raters, suggesting that the UI raters have developed a common understanding of the domain being rated and of how to rate it; that is, they have created a common image of the ideal candidate. There is no prearranged meeting to discuss the definitions of the criteria, the probing questions to ask, or what an ideal response would be. It is likely that the UI members (experienced faculty members and senior residents) have developed their "ideal candidate" image based on the residents currently in the program who perform well and are well integrated. An additional potential explanation relates to the a priori review of applicants' files, which might have caused UI members to develop an overall opinion of each candidate based on elements outside of the interviews, inducing a higher interitem consistency. There is an interval of 2 months between the file review and the interviews. On the interview day, raters have access to a brief summary of the candidate's file listing the completed rotations, but not including the reference letters.

The UI criteria differ in number and content from the dimensions used in the SI; although this difference creates an additional source of variability, the criteria were not altered so as to not manipulate the traditional UI.

When assessing the domains deemed essential to performing well, UIs are not more reliable ($E\rho^2$, 0.50 and 0.46) than the SI. The main VC for both panels still results from the interrater interaction. Interitem consistency remains high, with Cronbach α from 0.75 to 0.97. Factor analysis again uncovers only 1 factor. The poor interrater reliability suggests either that this time interviewers have different definitions of the dimensions being rated, or that the rating scheme differs between interviewers on the same dimension (a great response for one rater is rated as poor by a fellow rater). This is an important finding because the 4 domains listed have been identified after a rigorous job analysis as the essential ones to seek in applicants to our program. The strength of the SI rests in its ability to discriminate between dimensions.

Our results support the continuous use of SIs despite their labor-intensive development. The optimization of the overall reliability of our SI requires an improvement of its interitem reliability. This could be accomplished by increasing the number of scenarios per dimension, to reach saturation, and increasing the duration of the interviews. Alternatively, 4 SIs could be implemented, each assessing only 1 dimension, with several scenarios per dimension. The high interrater reliability suggests that a panel could be composed of only 2 interviewers, allowing more panels to be established with the same number of faculty members.

The SI assessed in this study was designed for applicants to an EM program. Although the tool itself might be too specific for use by other specialties, the study results are

applicable to all programs, namely, (1) SIs are worth developing, (2) a well-structured interview yields a high interrater reliability, (3) SIs discriminate well between various dimensions, and (4) each domain should be assessed by several items/scenarios to reach saturation and increase the interitem reliability, and consequently the SI overall reliability.

Our study has several limitations. Because our SI assesses different numbers of domains and domain contents than the UIs, differences in reliability should be weighed against potential differences in validity. The purpose of the study was to determine the performance of an SI compared with that of the current interview tool, hence the necessity to keep intact the traditionally used interview. In an attempt to overcome this limitation, the UI members were asked to rate candidates on the same 4 dimensions marked by the SI in addition to the traditional interview dimensions.

The UIs were by definition not scripted; the types of questions asked, the follow-up questions, the interviewers' tone, and the possibility for the interviewers to consult each other between candidates all could have influenced the ratings. These factors would in effect increase the interrater reliability; our results show poor UI interrater reliability despite all of these possible confounders.

Bootstrap estimates of standard errors have been noted to be biased in generalizability theory. The small sample size of raters and items imposes further limitations on the bootstrap methods. Thus, the standard errors provided for the reliability coefficients should be interpreted as rough approximations.

The small number of residency positions for any specialty in Canada intrinsically limits the number of applicants that will be granted interviews. In addition, the interview process at our institution only allows for 20-minute interviews, restricting the number of scenarios and domains with which to assess candidates. A greater number of scenarios would have better saturated the domains studied and yielded higher overall reliability.

Conclusions

When tested on the same pool of applicants, our SI did not achieve greater overall reliability than the UIs. This was a consequence of the multidimensionality of the SI inducing poor interitem reliability; the consistency between raters was actually higher than that of the UIs. The SI was successful at discriminating between several dimensions, whereas the UIs consistently assessed only one. When prompted to assess the dimensions identified as essential for residents to perform, unstructured panels persisted in assessing a single dimension, this time with poor overall reliability due to a reduced between-candidate variance. Constraints in the duration of the interview process limited the possible saturation of the domains assessed by the SI. Allowing the structured tool to more deeply assess each dimension by increasing the number of questions per dimension would lead to higher interrater reliability. Options include one long SI assessing all dimensions or several shorter interviews each assessing one dimension.

References

1 Galazka SS, Kikano GE, Zyzanski S. Methods of recruiting and selecting residents for U.S. family practice residencies. *Acad Med*. 1994;69(4):304–306.
2 Provan JL, Cuttress L. Preferences of program directors for evaluation of candidates for postgraduate training. *CMAJ*. 1995;153(7):919–923.
3 Wagoner NE, Suriano JR, Stoner JA. Factors used by program directors to select residents. *J Med Educ*. 1986;61(1):10–21.
4 Baker HG, Spier MS. The employment interview: guaranteed improvement in reliability. *Public Pers Manage*. 1990;19(1):85–90.
5 Garden FH, Smith BS. Criteria for selection of physical medicine and rehabilitation residents: a survey of current practices and suggested changes. *Am J Phys Med Rehab*. 1989;68(3):123–127.
6 Taylor CA, Weinstein L, Mayhew HE. The process of resident selection: a view from the residency director's desk. *Obstet Gynecol*. 1995;85(2):299–303.
7 Bernstein ADM, Jazrawi L, Elbeshbeshy B, Valle CJD, Zuckerman J. Orthopaedic resident-selection criteria. *J Bone Joint Surg Am*. 2002;84-A(11):2090–2096.
8 Janis JEM, Hatef DAM. Resident selection protocols in plastic surgery: a national survey of plastic surgery program directors. *Plast Reconstr Surg*. 2008;122(6):1929–1939.
9 Bajaj GM, Carmichael KDM. What attributes are necessary to be selected for an orthopaedic surgery residency position: perceptions of faculty and residents. *South Med J*. 2004;97(12):1179–1185.
10 Nallasamy S, Uhler T, Nallasamy N, Tapino PJ, Volpe NJ. Ophthalmology resident selection: current trends in selection criteria and improving the process. *Ophthalmology*. 2010;117(5):1041–1047.
11 LaGrasso JRM, Kennedy DAM, Hoehn JGM, Ashruf SMD, Przybyla AMM. Selection criteria for the integrated model of plastic surgery residency. *Plast Reconstr Surg*. 2008;121(3):121e–125e.
12 Hermelin E, Robertson IT. A critique and standardization of meta-analytic validity coefficients in personnel selection. *J Occup Organ Psychol*. 2001;74(3):253–277.
13 Patrick LE, Altmaier EM, Kuperman S, Ugolini K. A structured interview for medical school admission, phase 1: initial procedures and results. *Acad Med*. 2001;76(1):66–71.
14 Altmaier EM, Smith WL, O'Halloran CM, Franken EA Jr. The predictive utility of behavior-based interviewing compared with traditional interviewing in the selection of radiology residents. *Invest Radiol*. 1992;27(5):385–389.
15 Latham GP, Saari LM, Pursell ED, Campion MA. The situational interview. *J Appl Psychol*. 1980;65(4):422–427.
16 Campion MA, Palmer DK, Campion JE. A review of structure in the selection interview. *Pers Psychol*. 1997;50(3):655–702.
17 Campion MA, Pursell ED, Brown BK. Structured interviewing: raising the psychometric properties of the employment interview. *Pers Psychol*. 1988;41(1):25–42.
18 Olawaiye A, Yeh J, Withiam-Leitch M. Resident selection process and prediction of clinical performance in an obstetrics and gynecology program. *Teach Learn Med*. 2006;18(4):310–315.
19 Brothers TE, Wetherholt S. Importance of the faculty interview during the resident application process. *J Surg Educ*. 2011;64(6):378–385.
20 Spitzer AB, Gage MJ, Looze CA, Walsh MP, Zuckerman JD, Egol KA. Factors associated with successful performance in an orthopaedic surgery residency. *J Bone Joint Surg Am*. 2009;91(11):2750–2755.
21 Dubovsky SLM, Gendel MHM, Dubovsky ANM, Levin RP, Rosse JP, House RMD. Can admissions interviews predict performance in residency? *Acad Psychiatry*. 2008;32(6):498–503.
22 Metro DG, Talarico JFD, Patel RM, Wetmore AL. The resident application process and its correlation to future performance as a resident. *Anesth Analg*. 2005;100(2):502–505.
23 Blouin D, Dagnone JD. Performance criteria for emergency medicine residents: a job analysis. *CJEM*. 2008;10(6):539–544.

24 Blouin D. Reliability of a structured interview for admission to an emergency medicine residency program. *Teach Learn Med*. 2010;22(4):246–250.
25 Flanagan JC. The critical incident technique. *Psychol Bull*. 1954;51(4):327–358.
26 The Royal College of Physicians and Surgeons of Canada. Information by specialty or subspecialty. http://rcpsc.medical.org/information/index.php. Accessed on May 11, 2010.
27 Brennan RL. Performance assessments from the perspective of generalizability theory. *Appl Psychol Meas*. 2000;24(4):339–353.