Getting on the Same Page: The Impact of Interviewer Education and Structured Interviews on Interrater Agreement in Residency Interviews

Aimee K. Gardner, PhD, [*,†] Paula Costa, PhD, [†] and Ross E. Willis, PhD [‡]

[*]Baylor College of Medicine, Houston, Texas; [†]SurgWise Consulting, Houston, Texas; and [‡]University of Texas Health San Antonio, San Antonio, Texas

INTRODUCTION: We explored the impact of implementing structured interviews and associated interviewer education on interrater agreement within a large academic residency program.

METHODS: Faculty and senior resident interviewers from a large academic residency program participated in a 3-hour structured interview course. Before and after the course, participants completed a 15-item assessment pertaining to the characteristics, logistics, and guidelines associated with structured interviews. Along with interviewer training, interview day logistics also changed from an unstructured format (no specific questions, one overall 1-9 rating scale) to a structured interview format, including incorporation of behavioral-based competency questions that would be asked of every applicant and behavioral anchored rating scales (1-10; 10 = highest). Interrater agreement was assessed via intraclass correlation coefficients (ICC1) for the 2 years before and 2 years after incorporation of the structured interview format.

RESULTS: A total of 45 faculty and resident interviewers participated in the course in 2018. Participant knowledge significantly increased from an average of 36% to 79% after the course ($p < 0.01$). Prior to the intervention, overall interrater agreement was "poor" to "fair," with an ICC1 of 0.51 in 2016 and 0.49 in 2017. After the structured interview intervention, overall agreement increased to the "good" level with an ICC1 of 0.71 in 2018 and 0.66 in 2019. The proportion of applicants who received interview scores with at least 2 ratings more than 2 points apart significantly decreased from 59% to 47% after the intervention ($p < 0.01$).

CONCLUSIONS: Incorporating an interviewer educational session and a structured interview format into residency selection can help increase agreement in ratings between interviewers. However, these data suggest that ongoing refresher trainings may be needed to maintain acceptable levels of interrater agreement. (J Surg Ed 79: e12–e16. © 2022 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: interview, interviewer training, interrater agreement, structured interview

COMPETENCIES: Systems-Based Practice, Practice-Based Learning and Improvement

INTRODUCTION

In medical education settings, unstructured interviews are the norm. A survey of surgery residency programs revealed that only 5% of programs incorporate any form of structured questions into their interview process.[1] Instead, programs most commonly conduct unstructured interviews, in which applicants rotate through a series of interview rooms and are asked a variety of questions from faculty interviewers about their interest in the specialty, knowledge of the program, specifics about their application, and anything else intended to develop rapport with the candidate. Faculty may even have their "favorite" questions to put applicants on the spot or try to gauge their ability to think on their feet. Sometimes, programs proactively organize these rooms by "theme," and require faculty to ask questions according to a certain topic or competency, such as leadership, grit, or problem solving. However, even this practice is rare.

Despite the fact that these informal conversations may be well-received by applicants and may help in

Podium presentation at the Association of Program Directors in Surgery (APDS) Annual Meeting, May 2-5, 2022. San Antonio, TX.

Correspondence: Inquiries to Aimee K. Gardner, PhD, One Baylor Plaza, Houston, Texas 77030; e-mail: aimee.gardner@bcm.edu

  1931-7204/$30.00
https://doi.org/10.1016/j.jsurg.2022.05.013

developing rapport,[2,3] interviewers are often able to obtain little usable information from them. Unstructured interviews limit the ability to gather specific, competency-based data on each applicant, create difficulty in comparing candidates along the same dimensions, and do not ensure that rating forms, if they exist at all, are being used in the same way among interviewers. The literature supports these limitations, showing that unstructured interviews can lead interviewers to focus on irrelevant information and increase susceptibility to biases,[4,5] are highly unreliable,[6,7] are poor predictors of job performance,[5,6,8] and can actually *hurt* predictive accuracy compared to not even interviewing at all.[9,10]

The alternative is to adopt a structured interview. Structured interviews have 4 key characteristics:[11] The first, is that all questions are created prior to the interview, and are based upon a thorough job analysis—a rigorous, multi-method competency modeling process to help organizations identify key competencies required for success in their program. These data are used to develop program-specific interview questions and rating forms. Structured interviews also require that all candidates are asked the same exact questions, and in the same order to provide an equitable opportunity for applicants and reduce any primacy, recency, or contrast effects. Finally, structured interviews require that faculty are trained not only on how to conduct interviews to maximize utility and minimize bias, but that they are also trained to use the competency rating forms in the same way.

As a result of this structure, these types of interviews have a strong evidence-base behind them. They demonstrate higher levels of reliability between raters,[6,12] are better able to predict later job performance,[5,6,8] and minimize opportunities for racial and gender bias to emerge.[13] While less robustly researched in medical education settings, structured interviews have been shown to reduce bias during fellowship interviews.[14] Importantly, structured interviews are also more efficient. Studies have shown that it would take 3 to 4 unstructured interviews to reach the same validity levels as just one structured interview conducted by one interview.[15] In a field that conducts a high number of interviews with faculty who have busy clinical schedules, the value of this latter finding cannot be ignored. In summary, the structure and standardization embedded within structured interviews is important from the validity, reliability, fairness, and practicality perspectives. For all of these reasons, structured interview meet best practice and legal standards for an assessment method and are the recommended approach for residency interviews by the Association of American Medical Colleges.[16]

Unfortunately, there is a dearth of research within the surgical literature documenting the efficacy of a structured interview intervention on interviewer knowledge and interviewer agreement within a residency program. We explore the impact of implementing structured interviews and associated interviewer training program on changes in interviewer knowledge and interrater agreement within a large academic residency program.

TABLE 1. Course Components and Delivery Methods

Course Topic	Delivery Method				
	Didactics	***Video or Audio-based Case Review***	***Small Group Discussion***	***Role Play***	***Other Active Learning Strategies***
Part I: Background					
Structured interview basics	x		x		
Question development	x		x		
Biases in interviews		x	x		
Part II: Asking Questions					
Getting complete responses		x	x		
Types of interviewing questions	x		x		
Unacceptable and illegal questions		x	x	x	x
Taking notes		x			x
Part III: Assessment					
Assigning ratings		x	x	x	x
Motivational fit		x	x		
Integrating data	x		x		
Part IV: Putting it All Together					
Interview day basics	x				
Review		x	x		

METHOD

Training Intervention

Faculty and senior (PGY3-5) resident interviewers from a large academic residency program participated in a 3-hour structured interview course called InterviewWise previously described in the literature.[17] The course was led by 2 Industrial-Organizational Psychologists and consisted of numerous experiential learning formats (discussion, audience polling, audio and video vignettes) to cover concepts related to structured interview basics, common interviewer mistakes, biases, getting complete responses, utilizing probing questions, and behavioral anchor rating scales (Table 1). The course also offered participants the opportunity to practice and gain group consensus on new structured interview questions and behavioral-anchored rating forms designed for the residency program.

Before the course, all interviewers were given a 15-item pre-test to complete that assessed pre-existing interviewing knowledge. The assessment utilized a variety of question formats (multiple choice, fill in the blank, etc.) and covered areas pertaining to structured interview characteristics, types of probing questions, understanding of behavioral questions, common interviewing mistakes, and identification of inappropriate and illegal questions. Example items include, "All questions asked to applicants in an interview should be related to the __" (multiple choice) and "Which bias describes when shared experiences or other similarities lead to the interviewer giving an applicant a more positive evaluation?" (fill in the blank). All interviewers were then given the same test after course completion to assess learning.

Interview Structure Intervention

Along with interviewer training, interview day logistics also changed from an unstructured format (no specific questions, 1 overall 1-9 rating scale) to a structured interview format. Prior to the intervention, each candidate was interviewed by 4 pairs of interviewers (i.e., up to 8 interviewers). Each interviewer pair consisted of 2 faculty or 2 PGY5 residents. Interviewer pairs completed one evaluation of each candidate (i.e., 4 evaluations per candidate).

The structured interview intervention incorporated both content-related and evaluation-related components: 1) incorporation of behavioral-based competency questions based on a comprehensive job analysis; 2) asking all candidates the same behavioral-based competency questions; 3) inclusion of behavioral anchored rating scales (1-10; 10 = highest); and 4) training interviewers as previously described. Number of interviewers per candidate and the length of interviews did not change. However, after the intervention, all interviewers completed individual evaluations (compared to one evaluation per room/pair in the previous system). Fewer candidates were interviewed each year after implementation of the structured interview process (2016: 168 candidates; 2017: 164 candidates; 2018: 133 candidates; 2019: 129 candidates) as the program implemented additional screening assessments to make more informed interview invitation decisions and because of increased reliance on the structured interview process each year.

Interview evaluations from the 2 years prior to the intervention and the 2 years after the interview were collected to identify any changes in interrater agreement. Data analyses were conducted using SPSS version 28. Paired samples t-tests were used to measure changes in knowledge on the pre and post course assessment. In addition to means and standard deviations, interrater agreement was assessed via intraclass correlation coefficients (ICC1) on overall agreement for the 2 years before and 2 years after incorporation of the structured interview format. As the assessment tools varied before and after the intervention (going from 1 universal rating to multiple competency-based ratings for each applicant), analysis was limited to changes in overall interrater agreement for the 2 time periods. This work was deemed quality improvement by the Institutional Review Board and thus not required to undergo review.

RESULTS

A total of 45 faculty (N = 15) and PGY3-5(N = 30) resident interviewers participated in the course in September 2018. All interviewers in 2018 and 2019 completed the course. Before the course, interviewers achieved an average of 36% correct on the pre-course assessment. Participants demonstrated significant increases in knowledge related to structured interviews after the course, with average post-course assessment performance of 79% correct ($p < 0.01$). There were no differences in structured interview knowledge changes between faculty and resident interviewers.

A total of 3971 interview evaluations (1143 pre-intervention; 2828 post-intervention) were available for these analyses. Prior to the structured interview intervention, overall interrater agreement was "weak," with an ICC1 of 0.51 (confidence interval [CI]: 0.16-0.33) in 2016 and 0.49 (CI: 0.12-0.30) in 2017. After the structured interview intervention, total overall agreement significantly increased to the "good" level with an ICC1 of 0.71 in 2018 (CI: 0.29-0.47) and 0.66 (CI: 0.23-0.44) in 2019 ($p < 0.001$).[18]

Prior to structured interviews, over half (59%) of all applicants received interview scores with at least 2

ratings more than 2 points apart across interviewers on the 9-point scale. After incorporation of structured interviews, only 43% of interview scores had at least 2 ratings over 2 points apart on the 10-point scale. These improvements were significant at the $p < 0.01$ level.

DISCUSSION

These data reveal that faculty and resident interviewers more than doubled their baseline knowledge after participating in a 3-hour structured interview training course designed specifically for surgeons. Average baseline scores of approximately 36% reveal that the average interviewer had not been adequately exposed to or trained on the basics of conducting structured interviews, common interviewer mistakes and biases, and the importance of avoiding inappropriate questions. Given these results combined with Hern et al.'s work demonstrating that the majority of applicants to general surgery are more likely than other specialty applicants to be asked inappropriate and illegal questions during residency interviews,[19] it is clear that opportunities to increase residency interviewer knowledge of structured interviews are needed.

Fortunately, participation in the course resulted in significant knowledge increases for interviewers. Of note, however, average post-course scores did not reach a uniform 100% across all interviewers. We attribute this variability to the varied participation and engagement among faculty, some of whom could only participate in a portion of the course and/or intermittently stepped out to answer phone calls or pages. Anecdotally, those who were most invested in the course, such as residency education team members and other educational leadership, most often achieved 100% on the post-course assessment, suggesting that engagement may have indeed played a role. Since this study was conducted, the course has moved to an online format, accommodating turbulent surgeon schedules and allowing program leadership to require interviewers to reach uniform proficiency prior to interviewing.

We sought to measure the ultimate impact of our training by examining interrater agreement of interview ratings before and after the intervention. Overall, we found that agreement in interview ratings across interviewers were in the "weak" range prior to the interview, and significantly increased into the "good" level immediately after implementation. While improvement was still shown in the second year after intervention, the slight drop in agreement does suggest that providing refresher training to interviewers who previously participated and training any new interviewers (e.g., additional incoming senior residents, faculty who did not previously participate) is a fruitful endeavor. In fact, since review of these data, the program has begun distributing the online course to all faculty preceding interviews. These results also suggest that programs should not expect perfect agreement between interviewers even after training interviewers and increasing structure.

Of note, these data do not include any data after the implementation of virtual interviews. However, the need for interviewer training may be even more important in an era of virtual interviews. With few opportunities to meet candidates and develop rapport, interviewers may feel even more obliged to "get to know" candidates by asking questions related to their personal life, such as marital status, place of origin, or family plans. Unfortunately, inquiries used to develop rapport can often fall into the category of inappropriate or illegal questions. Although interviewers may have the best of intentions, these questions may be perceived as discriminatory, leading applicants to rank the program lower or not at all.[19] They may also open the door for potential litigation for the program. Enhancing interviewer knowledge and skills can also aid faculty in acknowledging and avoiding potential biases that may be just as, if not more so, present during virtual interviews, such as the just-like-me bias, halo/horn bias, attractiveness bias, uniqueness effect, and contrast effects.[20] Thus, programs will be wise to implement some form of interviewer training prior to ensure both maximum utility and minimal construct irrelevant variance.

Of course, this study is not without its limitations. First, although we were able to double baseline knowledge regarding how to conduct fair and structured interviewers, we have no way to know the extent to which faculty interviewers actually retained or implemented these new tactics into the interviews they conducted over the course of residency selection season. For example, given that the pre and post course knowledge test occurred just a few hours apart, we are unable to ascertain retention or future application. We measure interrater agreement as a proxy for assessing changes in behavior, but this is obviously not a comprehensive or purely direct link. Similarly, not all interviewers participated in the training, likely minimizing the true positive impact of the intervention. Further, our evaluation of interrater agreement is limited to only the common evaluations completed across interviewers, and thus further exploration of specific competency questions asked by a single interviewer could not be included. Finally, our methodology does not allow us to fully explore if the increase in interrater agreement was a result of the training, enhanced structure, or both. Future work could dissect what aspect of this intervention was most impactful to better our understanding.

CONCLUSION

This multi-year study reveals that there is ample opportunity to improve interviewer knowledge and skills related to conducting structured interviews. Participating in a 3-hour course can improve faculty competency in this area. Programs implementing structured interviews and training faculty on how to conduct them can also realize greater agreement across interviewers for up to 2 years later.

CONFLICT OF INTEREST

Paula Costa, PhD is employed by SurgWise Consulting. Aimee K. Gardner, PhD is part owner of SurgWise Consulting. Ross Willis, PhD has no relevant conflicts to report.

REFERENCES

1. Kim RH, Gilbert T, Suh S, Miller JK, Eggerstedt JM. General surgery residency interviews: are we following best practices? *Am J Surg*. 2016;211:476–481.
2. Latham GP, Finnegan BJ. Pereceived practicality of unstructured, patterned, and situational interviews. Schuler H, Farr JL, Smith M, editors. Series in Applied Psychology. Personnel Selection and Assessment: Individual and Organizational Perspectives, Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc; 1993:41–55.
3. Schuler H. Social validity of selection situations: a concept and some empirical results. Schuler H, Farr JL, Smith M, editors. Series in Applied Psychology. Personnel Selection and Assessment: Individual and Organizational Perspectives, Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc; 1993:11–26.
4. Nisbett RE, Zukier H, Lemley RE. The dilution effect: nondiagnostic information weakens the implications of diagnostic information. *Cog Psychol*. 1981;13:248–277.
5. Wiesner WH, Cronshaw SF. A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *J Occup Psychol*. 1988;61:275–290.
6. McDaniel MA, Whetzel DL, Schmidt FL, Maurer SD. The validity of employment interviews: a comprehensive review and meta-analysis. *J Appl Psychol*. 1994;79:599–616.
7. Viswesvaran C, Ones DS, Schmidt FL. Comparative analysis of the reliability of job performance ratings. *J Appl Psychol*. 1996;81:557–574.
8. Huffcutt AI, Arthur W. Hunter and Hunter (1984) revisited: interview validity for entry-level jobs. *J Appl Psychol*. 1994;79:184–190.
9. Dana J, Dawes R, Peterson N. Belief in the unstructured interview: the persistence of an illusion. *Judg Decision Making*. 2013;8:512–520.
10. Kausel EE, Culbertson SS, Madrid H. Overconfidence in personnel selection: When and why unstructured interview information can hurt hiring dcisions. *Org Behav Human Decision Processes*. 2016;137:27–44.
11. Campion MA, Palmer DK, Campion JE. A review of structure in the selection interview. *Personnel Psychol*. 1997;50:655–702.
12. Conway JM, Jako RA, Goodman DF. A meta-analysis of interrater and internal consistency reliability of selection interviews. *J Appl Psychol*. 1995;80:565–579.
13. McCarthy JM, Van Iddekinge CH, Campion MA. Are highly structured job interviews resistant to demographic similarity effects? *Pers Psychol*. 2010: 325–359.
14. Langhan ML, Goldman MP, Tiyyagura G. Can behavior-based interviews reduce bias in fellowship applicant assessment? *Acad Peds*. 2022;3:478–485.
15. Schmidt FL, Zimmerman RD. A counterintuitive hypothesis about employment interview validity and some supporting evidence. *J Appl Psychol*. 2004;89:553–561.
16. AAMC. Best Practices for Conducting Residency Program Interviews. Available at: https://www.aamc.org/media/44746/download Accessed March 23, 2022.
17. Gardner AK, Donofrio BC, Dunkin BJ. Can we get faculty interviewers on the same page? An examination of a structured interview course for surgeons. *J Surg Educ*. 2018;75:72–77.
18. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med*. 2012;22:276–282.
19. Hern HG, Alter HJ, Wills CP, Snoey ER, Simon BC. How prevalent are potentially illegal questions during residency interviews? *Acad Med*. 2013;88:1116–1121.
20. Roulin R. The Psychology of Job Interviews. New York, NY: Routledge; 2017.