The Panel Interview: A Review of Empirical Research and Guidelines for Practice

Marlene Dixon
Sheng Wang
Jennifer Calvin
Brian Dineen
Edward Tomlinson

After over 50 years of research, the panel interview remains an important yet controversial tool for personnel selection. Previous narrative and meta-analytic reviews have yielded conflicting results concerning its reliability and predictive validity. Furthermore, no review has focused exclusively on the panel interview. By examining the features and psychometric property of the panel interview, we can not only add to the scholarly literature but also determine important, research-based applications for the practitioner. We have derived an eight-step panel interview procedure from previous research. Utilizing this procedure as an organizing framework, this review highlights various features of the panel interview including: setting, structure and scoring anchors, question type, training, and rating combination method. Each of these features is discussed in terms of interview trends and in relation to reliability and validity. Practical implications and directions for future research also are addressed.

After nearly 50 years of research, examining the overall utility of the panel interview continues to be important, especially given its wide use in the public sector and apparent growing use in the private sector. For practitioners, considerations such as cost, face validity, adverse impact, and legal defensibility are important when choosing tools for selecting employees.[1] For example, practitioners have argued that the personnel costs (in terms of hours spent interviewing) are greater for panel interviews than those of the individual interview. These costs, however, might be offset by the greater predictive and face validity of panel interviews.[2] In addition, panel interviews may increase buy-in among incumbents regarding the ultimate selection decision.[3] By examining the features and psychometric properties of panel interviews, we can determine important, research-based implications for the practitioner.

One may argue that after 50 years, we have learned all we can about panel interviews. However, based on past meta-analyses, narrative reviews, and empirical research, it is clear that much investigation remains. In spite of the intuitive appeal

prompting the use of panel interviews, research findings have generally yielded equivocal results.[4] While some meta-analyses have determined panel interview validity to be as high as .44,[5] others have found it to be a dismal –.04.[6] In fact, one meta-analysis reported different results depending on the studies utilized.[7]

Furthermore, while a number of narrative reviews have been conducted on interviews as a selection device, it is quite surprising that no review, to date, has focused exclusively on the panel interview. Most of the past research has focused either on the individual versus the panel interview[8] or on the interview versus other selection devices.[9] Thus, while others have offered advice to the practitioner concerning the use of the panel interview[10] they have based their suggestions on reviews of mixed types of interviews or on an incomplete review of the literature.

The absence of a literature review focused solely on panel interviews, the disagreement between meta-analytic studies, and the lack of a systematic review of research to substantiate recommendations offered, leaves many questions unanswered for the scientist and practitioner regarding the panel interview. The purpose of our review here is to: (a) describe in detail the important elements of panel interviews with particular focus on those elements that have been suggested as moderators of reliability and validity, (b) review the literature on panel interviews, noting discrepancies and potential explanations for them, and (c) propose practical implications and directions for future research. The reviewed research relies primarily on meta-analyses and other published, peer-reviewed articles. However, some dissertations, government documents, or presentations have been included if they provide particularly salient points relevant to the panel interview literature.

Definition and Background

A panel interview, also known as a board interview, is defined as an interview conducted by a team of interviewers (usually two to three), who interview the candidate simultaneously, then combine their ratings into a final panel score.[11] The panel interview is in contrast to the individual interview, whereby one interviewer rates one candidate, or the serial interview, whereby multiple interviewers assess a single candidate, but they do so sequentially instead of simultaneously. The panel interview process has evolved considerably over the past fifty years from "informal discussions" between candidates and the panel[12] to highly structured situational and behavioral assessments based on strict job analysis, complete with interviewer training and scoring anchors.[13]

Campion et al. claimed, "a disproportionately large number of studies on panels are in the public sector."[14] Many of them were utilized in police and military settings.[15] Though panel interviews have also been studied in other civil service settings.[16] One of the earliest studies was for selection of patrolmen in 1947.[17]

All studies on panel interviews prior to 1980 were conducted exclusively in the public sector.[18] The earliest panel interview using private settings was conducted in 1980.[19] Latham et al. reported on three studies in their article; two of them examined

sawmill workers using incumbents rather than applicants as their subjects, and the third study was conducted among applicants for entry-level pulp mill positions.[20]

Since 1980, four published studies on panel interviews used public settings, while five used private settings.[21] While research in the public sector continues, these numbers may be evidence that a trend exists toward adding the private sector in panel research. It also may reflect a trend of more private organizations using panel interviews for selection.[22]

Conflicting Results

Meta-analyses continue to produce conflicting results regarding the panel interview, depending on the sample used and the moderators chosen.[23] The most recent meta-analysis by Huffcutt and Woehr indicates that research regarding the use of panel interviews as opposed to individual interviews has been inconclusive.[24] For example, according to Wiesner and Cronshaw,[25] the predictive validities for all individual and panel—both interviews structured and unstructured—were the same (.44), while McDaniel et al.[26] reported that individual interviews (.43) were more predictively valid than panel interviews (.32). Perhaps even more interesting is that although both Wiesner and Cronshaw and McDaniel et al. examined interview format by introducing structure as a moderating variable, they did not obtain similar results. Wiesner and Cronshaw indicated that unstructured panel interviews had significantly higher validity than unstructured individual interviews (.37 and .20 respectively), while McDaniel et al. found no significant difference between the two types (unstructured individual was .34 and unstructured panel was .33). These same two studies also found that when the interviews were highly structured, the individual interview had similar or higher validity than panel interviews.

Huffcutt and Woehr's meta-analysis of four possible influences on the validity of interviews found that interview format had a very small, statistically non-significant, and negative correlation with validity even after correction for sampling error (–.05).[27] They suggested that using a panel interview could have a detrimental effect on validity. However, much of the research analyzed in all of these meta-analyses included studies done in the laboratory not the field setting. Furthermore, significant differences in the data sets were included in the various meta-analyses.

Using different data sets and different definitions seems prevalent and problematic throughout the reviews. Wiesner and Cronshaw, for example, used a very small sample of board (a.k.a. panel) interviews, as not many suitable studies could be found on that topic for inclusion in their meta-analysis.[28] Marchese and Muchinsky reported that the "relationship between the number of interviewers and the validity of the interview was non-significant."[29] However, in the Marchese and Muchinsky study, no distinction was made between panel interviews and serial one-on-one interviews. Further, they cite studies where duplicate meta-analyses found different results. These differences in results may be attributable to the judgment calls required during the process

of locating relevant studies, of setting the criteria for including studies in meta-analyses, and of assigning meaning to data from the studies selected.

A number of moderators that might account for the differences in meta-analytic results also have been suggested. First, interview structure has been discussed as a possibility.[30] However, as stated earlier, structure did not account for the discrepancies. Wiesner and Cronshaw proposed that the method of combining scores might be a moderator.[31] They reported that consensus of the panel members, rather than statistical averaging of the panel's ratings, appeared to produce higher predictive validity. However, they also based their conclusions on a rather small sample size and suggested interpreting them with caution. Huffcutt and Woehr found that interviewer training was the strongest moderator of validity, greater than note taking, structure, or use of a panel interview.[32]

In summary, previous narrative reviews and meta-analyses have left many questions unanswered regarding the panel interview. Literature reviews comparing panel interviews to individual interviews or other selection tools have failed to investigate differences within a single tool. It may be that specific features of the panel interview are critical to its validity and overall utility as a selection device. Meta-analyses have produced conflicting results regarding reliability and validity. This paper seeks to explain differences in these results by highlighting important features of the panel interview. Furthermore, we seek to use the information generated in this review to defend and expand previous practical suggestions. By combining previous practical suggestions,[33] we have developed an eight-step framework for conducting panel interviews (see Figure 1). For each step, we define and highlight features of panel interviews, demonstrating trends in their design and implementation. Next, we review in detail the findings of past empirical research and meta-analyses, with the intent of explaining conflicting results regarding variables such as setting, structure and scoring anchors, question type, number of interviewers, training, and combination methods. Table 1 shows a comparison of these variables across studies. Finally, we propose ideas for future applied research along with research-based guidelines for the practitioner.

Figure 1. Summary of Recommendations

1. Perform job analysis
2. Develop questions-behavioral, situational, or both
3. Develop scoring anchors with at least high, medium, and low sample responses and keywords
4. Select panel members-three to six individuals, mixed membership
5. Train panel
6. Conduct interviews
7. Evaluate candidates, reaching a consensus on ranking or ratings
8. Evaluate selection decisions based on subsequent employee performance

Table 1. Features of the Panel Interview

Author/Year:	Glaser, Schwartz, & Flanagan, 1958	Flynn & Peterson, 1972
Setting:	Civilian supervisors at military depots	Police recruits in training
N:	80	39
Structure:	Unstructured	Not reported
Type:	Behavioral	Behavioral
Panel Size:	3	Not reported
Training:	Not reported	Not reported
Consensus vs. average:	Average	Consensus
Scoring anchors:	None	Not reported
Inter-rater reliability:	Depot A: r=.77; Depot B: r=.47 Depot C: r=.65	Not reported
Predictive validity:	Performance on three summed criteria r=.12 (combined samples)	Score on final exam r=.345. Incremental validity (after training and experience and Public Personnel Exam) r=.03

Author/Year:	Gardner & Williams, 1973	Landy, 1976
Setting:	British Royal Navy	Police jobs
N:	269	399 interviewed, 150 hired
Structure:	Not reported	Structured
Type:	Not specified	Behavioral
Panel Size:	7	3
Training:	Not reported	Not reported
Consensus vs. average:	Consensus	Average
Scoring anchors:	Not reported	None
Inter-rater reliability:	Separate dimensions r=.81–.98; Not reported	Overall recommendation r=.94
Predictive validity:	Early training performance r=.34–.37; Specialist performance r=.30; Time to promotion r= –.25–.31	Overall performance n.s.; Technical Competence r=.26–.33; Demeanor r=.29; Communication r=.34

Table 1. Features of the Panel Interview *(continued)*

Author/Year:	Anstey, 1977	Reynolds, 1979
Setting:	Civil service selection board	Police Officers
N:	301	67
Structure:	Not reported	Partially structured
Type:	Behavioral	Not specified
Panel Size:	Not reported	3
Training:	Not reported	Not reported
Consensus vs. average:	Consensus	Average
Scoring anchors:	Not reported	None
Inter-rater reliability:	Not reported	Individual dimensions r=.54–.66; Overall rating r=.90
Predictive validity:	Not reported	Not reported

Author/Year:	Latham, Saari, Pursell, & Campion, 1980	Borman, 1982
Setting:	Study 1&2: Sawmill workers; Study 3: Applicants to sawmill	Soldiers in recruiter training
N:	(1) 49 laborers; (2) 63 foremen; (3) 56 hires	57
Structure:	Structured	Not reported
Type:	Situational	Not reported
Panel Size:	2	2
Training:	Not reported	Not reported
Consensus vs. average:	Consensus	Consensus
Scoring anchors:	Yes, 5,3,1	Not reported
Inter-rater reliability:	Hourly workers r=.76; Foremen r=.79; Blacks r=.87; Females r=.82	Separate dimensions r=.44–.92; Median=.76; Overall r=.84
Predictive validity:	Laborers BOS r=.46; Foremen BOS r=.41; Job performance (12 months) r=.39 (females), r=.33 (blacks)	Performance on early exercises r=.07 (n.s.); Performance on later exercises r=.26 (n.s.)

Table 1. Features of the Panel Interview *(continued)*

Author/Year:	Reilly & Chao, 1982	Davey, 1984
Setting:	Review	Police jobs
N:	987	790 original; 121 completed
Structure:	Both	Structured
Type:	N/A	Not specified
Panel Size:	N/A	3
Training:	N/A	Full-day
Consensus vs. average:	N/A	Consensus
Scoring anchors:	N/A	Yes – high, medium, and low
Inter-rater reliability:	Not reported	r=.91=.99
Predictive validity:	R=.19	r=–.02–.79 for six different panels

Author/Year:	Latham & Saari, 1984	Campion, Pursell, & Brown, 1988
Setting:	(1) office clerks; (2) entry level in newsprint mill	Entry level employees in pulp and paper mill
N:	(1) 29; (2) 349	243 applicants, 149 hired
Structure:	Structured	
Type:	(1) Situational and behavioral (2) Situational	Behavioral
Panel Size:	2–3	3
Training:	Not reported	Not reported
Consensus vs. average:	Consensus	Does not specify
Scoring anchors:	Yes –5,3,1	Yes
Inter-rater reliability:	(1) Situational questions r=.81; Past experience r=.83; (2) r=.90	Overall rating r=.88
Predictive validity:	BOS x situational r=.39–.42; (1) BOS x past experience r=.14–.15 (n.s.); (2) BOS x situational r=.14	Supervisor rating of performance – uncorrected r=.34, corrected r=.56

Table 1. Features of the Panel Interview *(continued)*

Author/Year:	Wiesner & Cronshaw, 1988	Stohr-Gillmore, Stohr-Gillmore, & Kistler, 1990
Setting:	Meta-analysis	Count jail officers
N:	51.459	(1) 36; (2) 33
Structure:	Both	Structured
Type:	N/A	Situational
Panel Size:	N/A	3
Training:	N/A	Not reported
Consensus vs. average:	N/A	Consensus
Scoring anchors:	N/A	Yes – most effective, least effective
Inter-rater reliability:	Overall rating r=.85; Board was .07 higher than individual; Structured was .21 higher than unstructured	Hirees w/o situational interview r=.51–.68; Hirees w/ the situational interview r=.55–.71; Not hired r=.71–.75
Predictive validity:	All: P=.44 single; P=.44 panel; Unstructured: P=.20 single; P=.37 panel Structured: P=.63 single; P=.60 panel	On-the-job performance r=.35 (n.s.); Performance at one year r=.19; Performance after one year r=.39

Author/Year:	Lin, Dobbins, & Farh, 1992	Roth & Campion, 1992
Setting:	Custodians in large urban school district	Petroleum technicians
N:	(1) 1645; (2) 1160	3169 applicants; 934 tested; 177 hired
Structure:	Structured	Structured
Type:	Both	Not specified
Panel Size:	2	2
Training:	1 hour for conventional; 3 hours for situational	Two day
Consensus vs. average:	Consensus	Consensus
Scoring anchors:	Consensus – no; Situational – yes	No
Inter-rater reliability:	Conventional structured r=.99 Situational r=.99	Not reported
Predictive validity:	Not reported	Training r=n.s.; Performance r=.41; Promotion r=.20

Table 1. Features of the Panel Interview *(continued)*

Author/Year:	Green, Alter, & Carr, 1993	Marchese & Muchinsky, 1993
Setting:	State law enforcement agency	Meta-analysis
N:	32	31 studies, total not reported
Structure:	Structured	N/A
Type:	Behavioral	N/A
Panel Size:	3	N/A
Training:	8 hours	N/A
Consensus vs. average:	Consensus	N/A
Scoring anchors:	Very specific, with answers and word cues. 5,3,1, ratings	N/A
Inter-rater reliability:	Panel A: r=.72; Panel B: r=.57	Not reported
Predictive validity:	35 item measure, consensus r=.42, panel average r=.44; 8 item measure, consensus r=.40, average r=.41; overall corrected r=.81	All interviews r=.49; Single vs. panel r=–.20 (n.s.)

Author/Year:	McDaniel, Whetzel, Schmidt, & Maurer, 1994	Campion, Campion, & Hudson, 1994
Setting:	Meta-analysis	Southeastern pulp mill
N:	86,311	70
Structure:	N/A	Structured
Type:	N/A	Both
Panel Size:	N/A	2–3
Training:	N/A	One day
Consensus vs. average:	N/A	Average
Scoring anchors:	N/A	Yes – 5,3,1
Inter-rater reliability:	Not reported	r=.93
Predictive validity:	All interviews: P=.43 single, P=.32 panel; Unstructured: P=.34 single, P=.33 panel; Structured: P=.46 single, P=.38 panel	Performance overall r=,50 (uncorrected), r=.56 (corrected); Performance situational r=.39; Performance past behavior r=.51 (diff n.s.)

Table 1. Features of the Panel Interview *(continued)*

Author/Year:	Pulakos & Schmitt, 1995	Huffcutt & Woehr, 1999
Setting:	Large federal organization	Meta-analysis
N:	216 incumbents	18, 158
Structure:	Structured	N/A
Type:	Both	N/A
Panel Size:	3	N/A
Training:	One day	N/A
Consensus vs. average:	Consensus, but scores taken for reliability before consensus rating	N/A
Scoring anchors:	Yes – 1–7	N/A
Inter-rater reliability:	Experience-based r=.74–.86; Situational r=.76–.90	Not reported
Predictive validity:	Situational r= –.02 (n.s.); Experience-based r=.32	r= –.04

Review of Previous Panel Interview Literature

Step One: Conduct a Job Analysis

Description and Explanation

A job analysis can be defined as the collection of data about a job through observation, interviewing, questionnaires, chartings, and other means. The purpose of the job analysis is to provide factual data regarding the critical responsibilities, and the knowledge, skills, and abilities needed to perform the job. One of the most important reasons for basing the interview questions on a job analysis is to provide evidence of face and content validity, which is critical to defending interviewee selections in legal disputes.[34] The need for conducting a job analysis before interviewing has been so well established that it basically is a given not only for panel interviews but for all selection tests. All the panel interview studies reviewed, including those analyzing earliest, more informal panels utilized a job analysis for creating their interview questions.[35] Because there is essentially no variation in the use of this procedure, it will not be further reviewed, but simply re-iterated in regard to its centrality to panel interview procedures.

Step Two: Develop Questions

Description and Explanation

A potentially important factor in interview studies is the question type, which can be situational or behavioral.[36] Behavioral questions are based on the theory that past behavior predicts future behavior.[37] In a behavioral interview, candidates are asked to

describe their past experiences and relate them to the current position.[38] Situational questions are based on the theory that goals and intentions predict future behavior.[39] In a situational interview, the candidate is asked questions about hypothetical situations that might occur on a job and how they would handle them.

Both interview types have received considerable attention. Campion et al., for example, argues that question type may have an impact on perceived fairness, but outcomes regarding validity and reliability have been mixed.[40] For example, Latham and colleagues[41] have generally shown that situational questions have superior predictive validity, while more recently Campion et al.[42] and Pulakos and Schmitt[43] have argued for the superiority of behavioral questions. Both, however, tend to be higher in validity and reliability than other questions that may be ambiguous (e.g., self disclosure, goals, opinions) or inconsistent across candidates.[44] Furthermore, much recent attention has been devoted not only to question type but also to question wording and to appropriate settings for each type.[45]

Review

In the panel interview studies reviewed, reliabilities were quite similar for both question types. Three studies directly compared situational and behavioral interviews. Latham and Saari report only slightly different reliabilities in their study of office personnel. Using situational questions the reliability was .81; using past experience questions, the reliability was .83.[46] These reliabilities are similar to the earlier situational interview studies that showed reliabilities of .76–.82.[47] Pulakos and Schmitt also compared situational and experience-based interviews.[48] They reported situational reliabilities to be .76–.90 and experience-based as .74–.86. In contrast to Latham and Saari,[49] the situational interviews were connected with slightly higher reliability. Finally, Lin et al. compared situational to behavioral interviews and found reliabilities of .99 for both.[50] As noted earlier, however, their results are of little value in comparing interview types since they allowed interviewers to change their ratings before measuring reliability.

Other behavior-based studies report results consistent with these findings. Green et al.[51] and Stohr-Gillmore et al.[52] achieved fairly narrow reliabilities in their past-behavior interviews (.57–.72 and .55–.71 respectively). Landy reported very high reliabilities (.81–.98) with behavior-based interviews.[53] Although these three studies are more difficult to interpret because the interview types are not compared directly, the reliabilities for all of these studies are quite high. As mentioned earlier, however, the reliability may be a correlate of interview structure and the use of scoring anchors rather than question type. In structured interviews, both question types have demonstrated high reliability.

Conclusions regarding the validity of situational and behavior-based interviews are mixed. Across studies, the behavioral interviews show validities ranging from .12[54] to .44,[55] while the situational interviews range from –.02[56] to .46.[57] This general overview would suggest that the validity is similar, with situational interviews demonstrating a wider range. Further investigation of studies that directly compare the two question types may be more valuable.

Latham and colleagues were one of the first to systematically compare situational and behavioral panel interviews.[58] In the first part of their study, they compare the use of the two interview types in selecting office clerical personnel. In the second part, they utilize a situational interview to assess reliability and validity of situational interviews in a newsprint mill. In the first study, they utilized for both specific scoring anchors situational and behavioral questions. They compared the interview scores with a behavioral observation scale (BOS) completed by the interviewees' peers and supervisors at an unspecified time. Both the questions and the BOS were derived from a job analysis. For the experience questions, they reported non-significant validities of .14–.15. For the situational interviews, they reported validities of .39–.42 in study one and .14 in study two. They noted, however, that the results in study two might have been due to flawed methodology by the interviewers. Overall, they concluded that the situational interviews were superior.[59]

Campion et al. interviewed 70 pulp mill employees. They reported behavior-based interview validity to be .51, and situational interview validity to be .39. The difference, however, was statistically non-significant. In addition, they reported that the behavior-based interviews had incremental validity beyond the cognitive ability tests, while the situational ones did not.[60]

Pulakos and Schmitt also compared situational and behavioral interviews in a tightly controlled experimental situation. They argued that the inconsistency of results in previous research might have been due to question content, not simply presentation.[61] They argued that the Campion et al.[62] study was controlled for question type, but not for content. Furthermore, they suggested that order effects might have accounted for some variance.[63] Therefore, Pulakos and Schmitt asked each applicant only situational or only behavioral questions.[64] Second, they closely controlled the content of the questions to make them as similar as possible. Performance was rated by supervisors on a scale derived from the original job analysis (situational and behavioral questions also were derived from this source). The authors reported a –.02 validity for the situational interview and a .32 validity for the behavior-based. Their results seem to concur with Campion et al.[65] that behavioral interviews have superior predictive ability, when all other variables are constant. Although validity results seem conflicting, the controlled nature of the Campion et al. and especially the Pulakos and Schmitt study lend powerful evidence to support the superiority of behavior-based panel interviews over situational panel interviews.

Step Three: Develop Scoring Anchors

Description and Explanation

Another factor that may be related to reliability and validity is the use of scoring benchmarks. Previous literature strongly supports the high reliability of using benchmark answers derived from a job analysis.[66] In this type of interview, a thorough job analysis is conducted by a "committee" consisting of the research team, job incumbents, and supervisors/management. From the job analysis, situational and/or past behavior

questions are formulated. Then, this committee generates potential answers to the questions and rates them according to the quality of the answers. The interviews are constructed according to a structured format, where the same questions are asked of each applicant, usually in the same order. In the interview session, the interviewers are given both the questions and the benchmark answers (the interviewees are not shown the answers). They attempt to match the interviewee's response to one of their benchmark answers. Then, they give the interviewee the score corresponding to the answer. Final ratings are achieved either by averaging the individual scores,[67] or by coming to a final consensus decision.[68] They aid validity and reliability by making the same information salient to all interviewers and helping ensure that the information is interpreted and ranked consistently across the panel members.[69]

Review

Scoring anchors are a specific element of structure that may explain variance. In the present review, after removing the Davey[70] and Lin et al.[71] studies (because the reliability, as explained earlier, was obtained after raters were allowed to change their initial ratings), six studies utilized a benchmark scoring system, with reliabilities ranging from .55 to .90. The Green et al. (1993) study reported the lowest average inter-rater reliability. [72] The average reliability of their two panels was .65. These reliabilities were much lower than their pilot study (.87) and seem rather minimal considering the interviewers underwent eight hours of training prior to the study. Perhaps there were problems with the training session or with the instrument development. Still, these reliabilities are within the acceptable range. Stohr-Gillmore et al. also reported rather low reliabilities in their study of county jail correctional officers.[73] Based on summary recommendations, they reported ratings ranging from .55–.71. In this study, the interviewers were given benchmarks of "most effective" and "least effective." This scale may account for lower reliabilities because other studies used up to six different answers to score the responses.

The other studies reported much higher mean reliabilities with concurrent smaller ranges. Pulakos and Schmitt's experimental study reported reliabilities ranging from .74–.90.[74] These ratings were based on independent scores on six dimensions, showing rather high agreement among the raters on each of the dimensions. The interviewers were given a scale of 1–7 for their benchmark answers. Latham and colleagues' series of studies demonstrated high reliabilities with both situational and experience-based questions and, and utilized a 5-point benchmark scoring system (5= high, 3= medium, 1= low).[75] In their five studies, the reliabilities ranged from .76 to .90, with the highest correlations being reported for situational interviews with entry-level personnel.[76] Campion et al. reported similar results (.88) with a highly structured, five-point scale approach.[77]

Perhaps the available range of scores accounts for some of the difference in reliabilities, whereby five-point scales produce better results than two-point. Granted, this does not explain the Green et al. findings that utilized a 5-point scale and achieved rather low reliabilities compared to the other studies.[78] Another explanation is offered

by Latham and Saari, who suggested that low reliability indicates a lack of clarity in the benchmark scoring key.[79] Without further testing and sample questions, however this explanation cannot be substantiated. Still, it highlights the importance of pilot testing all interview questions and answers before administration and continually updating the measures based on reliability results.

In terms of validity, the use of scoring anchors proves rather unfruitful for adding insight into differences between the studies. Results of scoring anchor use nearly parallel those of structured interviews, regardless of anchors. In fact, Landy, who did not use any scoring anchors, reported similar validities (.26–.34) to any of the studies that utilized anchors.[80] Latham and colleagues, one of the biggest proponents of using scoring anchors, demonstrated high inter-rater reliabilities with the use of the anchors, but did not show that their use necessarily improved predictive validity.[81] They reported validities ranging from .14 to .46. Pulakos and Schmitt, likewise, reported high inter-rater reliabilities, which they attributed to the use of scoring anchors; but they found question type to be of greater importance in differentiating interview validities.[82]

In sum, well-designed scoring anchors seem to aid in improving inter-rater reliability in panel interviews. Their use, however, does not seem to demonstrate any patterns of improvement in interview validity, regardless of question type or performance criteria.

Step Four: Select Panel Members

Description and Explanation

In their meta-analysis of employment interview validity, Marchese & Muchinsky postulated that the number of interviewers might be related to interview validity.[83] Campion et al. suggested that panel interviews should provide higher reliability than individual interviewers.[84] Both of these generalizations, however, are based on reviewing one interviewer versus an unspecified number of multiple interviewers. In the studies reviewed, the range of interview panel size is quite small; all but one study utilized either two-or three-person panels. Gardner & Williams utilized a seven-person panel, but did not report reliability results.[85] Furthermore, while additional interviewers may enhance validity, the overall utility of the method may suffer with increasing panel size. In other words, costs, candidate intimidation, and/or administrative hassles may place practical limits on the panel size.[86] Thus, while the number of interviewers may be an important variable, it remains largely unexplored (and will not, consequently, be reviewed in the descriptive results).

Step Five: Train Panel Members

Description and Explanation

Huffcutt and Woehr suggested that training interviewers might enhance the validity of interviews. They argued, "Such training might establish a more systematic framework, thereby reducing differences among interviewers and increasing consistency

across applicants."[87] In their meta-analysis, training indeed had the strongest effect on validity of the four variables they investigated (which also included structure, note-taking, and format). Thus, training was investigated in this literature review as a potential source of variation in panel interview studies.

Review

Prior to 1992, only one study reported utilizing training for the interviewers.[88] After 1992, all the studies have employed some type of training. Lin et al. employed a 1-hour and 3-hour training sessions in each of the respective parts of their study.[89] Roth and Campion conducted a two-day training seminar.[90] Green et al.[91] and Pulakos and Schmitt[92] utilized one-day training sessions. In the earliest studies, training was utilized to instruct the interviewers on what to look for in an interview and how to identify the salient points.[93] In the more recent studies, however, training usually was employed to teach the raters how to use the anchoring scales.[94] There is little information available as to how the training was conducted. The studies simply report that a training session took place and that the purpose was to learn how to use the anchoring scale. Davey pointed out the importance of using the training session to weed out poor interviewers.[95] Later studies also reported problems with raters even after training.[96] The authors, however, did not seem to offer suggestions as to how the training could be improved.

Davey reported the most extensive information regarding the process of training in panel interviews.[97] The raters were put into panels of three to review video footage of interviewees. They then were scored and reviewed on the basis of their ratings and agreement. They discovered that the panels that performed well in the training (in terms of identifying salient information and coding it correctly) also had the highest validity averages of the six groups. Davey's study may be instructional in interviewer training in terms of alerting researchers as to who should be included as the study progresses. In other words, if the training session is not effective for some panels, those panels may need to be removed from the study.

Lin et al. employed 1-hour and 3-hour training sessions in each of the respective parts of their study.[98] Three other studies utilized a 1-day training session.[99] Green et al.[101] and Pulakos and Schmitt,[102] while both using a 1-day training session, reported quite different reliabilities and validities. Roth and Campion conducted a two-day training seminar, but did not report reliability.[100] Lin et al. reported a .99 reliability, after allowing raters to change their scores.[103] Likewise, Davey reported post-discussion reliabilities of .91–.99.[104] Thus, when utilizing post-discussion reliability, 1-hour and 1-day training sessions do not show appreciable differences.

Using pre-discussing ratings, Green et al.[105] reported reliabilities of .57 and .72 across the two panels, while Pulakos and Schmitt reported .74 and .90.[106] Both studies employed a structured interview with specific scoring anchors, thus to accounting for differences in reliability in terms of interview characteristics is difficult. Three possible reasons for the higher results reported in Pulakos and Schmitt are that 1) they provided superior training to their interviewers, 2) their interviewers were more experienced,

or 3) their anchoring scale was designed better.[107] Without more detailed information and testing, however, these questions are difficult to resolve.

In terms of validity, the three one-day training studies varied from –.02[108] to .79[109]. Roth and Campion, reported a range from .20 to .41[110] for their 2-day training. Thus, in contrast to Huffcutt and Woehr, these studies do not reveal a direct pattern regarding training and validity.[111] It may be that there are too few panel studies that report training to actually assess any significant trends. It may be that training is a significant moderator for all interviews (panel and individual), but not for panel only. It may be that training relates to reliability, which then relates to validity, whereas other interview features (such as structure, anchors, or type) may relate directly to validity.

In sum, training the panel members seems to be a valuable suggestion tool according to literature on individual and panel interviews. A review of panel interview literature, however, fails to reveal the same findings. The lack of information regarding personality traits of the panel members and how the training was conducted suggests more research is needed in these areas to further explicate the link between training and panel interview validity.

Step Six: Conduct the Interviews

A critical element of conducting interviews is to consistently administer the process to all candidates.[112] This process includes asking the same questions in the same order to all candidates. While this consistency is only one element of structure, it provides the basis of the structured approach. As suggested earlier, interview structure is particularly important for panel interviews. Without structure, interviewee responses are subject to different interpretations by each of the interviewers. The structure aids validity and reliability by making the same information salient to all interviewers and ensuring that it is interpreted and ranked consistently across the panel members.[113]

Campion et al. provides an excellent overview of how to actually conduct the interview, including the suggestion of asking the same questions in the same order. They suggest no prompting of candidates or follow-up questioning. They also suggest that the panel members take extensive notes so that they can recall answers from candidates who might have participated early in the process.[114]

Review

A number of authors have suggested that the more structured the interview the more reliable and valid it becomes.[115] Specifically, in terms of inter-rater reliability, Campion et al. argued that scores from multiple raters should be more reliable than single raters, especially when the raters use the same benchmark answers.[116] In other words, when the interviewers ask and hear responses to the same questions at the same time, their overall ratings of the interviewee should be more consistent than when different questions are asked and different responses are given (as in an individual or serial interview). Thus, panel interviews may be consistently higher in inter-rater reliability than single or serial interviews.[117] Wiesner and Cronshaw reported meta-analytic values of .85 for inter-rater reliability across 1,909 board interviews.[118] The

coefficients were .07 higher for board than for individual interviews, but showed the greatest differences between structured and unstructured panel interviews (.21). Thus, use of panels can be one way of enhancing structure. However, within the panel interview, more structure may provide additional gains in reliability and validity.

In the panel interview studies reviewed, ten used a structured format, two utilized an unstructured format, and four did not report on structure. At first glance, the structured interview reliabilities range from .55[119] to .99.[120] The reliabilities for unstructured interviews range from .47[121] to .99[122]. This simple comparison does not indicate a large difference between interview types, but a closer examination may reveal important discrepancies.

First, the inter-rater reliabilities reported in the Davey and Lin et al. studies are difficult compared to other studies since the raters utilized the post-discussion consensus method.[123] It is unfortunate that Lin et al. did not record prior-discussion reliabilities as well, especially since they compared unstructured to structured interviews (in terms of validity). If they had done so, their results could have been more easily compared with other studies. With these two studies removed, however, the range of reliability for structured interviews is .55–.90, with most studies reporting averages in the mid 70s to low 80s, while unstructured reliability ranges from .47–.77, with only one study utilizing this format.[124] It appears that the structured studies produce overall higher reliability. This generalization supports literature that suggests structure enhances reliability in interviews. While this suggestion initially was made in reference to panel versus individual interviews, these results demonstrate that it may hold across panel interviews as well.[125]

In relation to predictive validity, a number of authors have suggested that structured interviews should increase validity because they aid the interviewer in assessing only job-relevant information.[126] A quick scan of the structured studies indicates that predictive validities range from –.02[127] to .81.[128] Further investigation reveals that the majority of structured interview studies report uncorrected validities between .14[129] and .46[130] regardless of question type or validity criterion. This range, while wide, remains higher than the validity reported for unstructured interviews (.12).[131] Only when corrected validities are reported do figures reach above .50.[132] Furthermore, the differences between structured and unstructured interviews are consistent in the literature and in meta-analytic reviews. For example, Wiesner and Cronshaw calculated the corrected validity for structured interviews as .60, and that for unstructured as .37.[133] McDaniel et al. reported validity for structured as .38, and that for unstructured as .33.[134] The averages reported in the meta-analyses are similar to the range of corrected validities reviewed in this study.

In sum, it is probably accurate to conclude that structured interviews are more valid than unstructured. However, structure does not account for all the variance between the panel interview studies. The studies reviewed, while utilized structured interviews, differ as to the use of anchors and question type, which may account for additional variance. For example, Pulakos and Schmitt found a wide variance in validities depending upon question type (–.02–.32).[135] Latham and colleagues reported a

wide range of validity depending on the question type and performance criterion (.14–.46).[136] Stohr-Gillmore et al. also found a wide range of predictive ability that was not dependent on structure or type, but rather on performance criteria.[137] Thus, additional variables must be investigated in relation to the reliability and validity of panel interviews.

Step Seven: Evaluate the Candidates

Description and Explanation

There are essentially two approaches to combining the ratings by the members of the panel. One approach is to take an average of all ratings by interviewers and the other is to have interviewers discuss differences and reach a consensus.[138] Most of the studies used either the consensus method or a combination of statistical averages and consensus. One example is Pulakos and Schmitt, who utilized statistical averages (by taking each rater's score before the final discussion) to obtain inter-rater reliability.[139] They then had the panel decide on a consensus score, which became the final rating. Later, the performance scores (supervisor or peer ratings, BOS scales, later exercises) were compared to this final consensus score, not to the statistical average. This same method is utilized by Latham and colleagues in their series of studies.[140]

Certain pitfalls might be associated with each approach. For example, by taking a statistical average, members may not be exposed to all points of view of those on the panel, and individual biases may unduly influence ratings. On the other hand, consensus scores may suffer from "groupthink,"[141] whereby panel members become reluctant to challenge the groups' decision, in which case critical thinking deteriorates, or they suffer from social facilitation, whereby members change ratings to receive favorable evaluation from other group members, especially authority figure.[142] One method of combining ratings, however, has not demonstrated consistent superiority over the other,[143] and either appears acceptable in the panel interview literature.

Review

In terms of inter-rater reliability, four structured panel interview studies employed statistical averages and reported inter-rater reliability ranging from .78[144] to .98.[145] This narrow range reveals consistently high inter-rater reliability. The studies that utilize consensus ratings[146] reported a wider range of inter-rater reliability, from .51[147] to .90[148]. Overall, these ranges would suggest that statistical averages produce higher reliability than pre-discussion consensus scores.

Davey and Lin et al. calculated the inter-rater reliability using the independent ratings after discussion among interviewers.[149] Both of them show a remarkably high reliability result of above .91. Davey obtained a reliability range of .91 to .99 and Lin et al. reported r= .99. This result is not surprising, as interviewers were permitted to consult with each other, change their ratings if desired, and then rate the interviewees again. While this method may produce high reliability, the results may not reflect the quality of the instrument, but rather the panel's ability to come to consensus.

However, the validity of the panel interview does not appear related to use of either consensus or statistical combination. One reason is that most studies use the statistical average to calculate intraclass correlation (reliability) and then use the consensus score as the predictive measure. Therefore, it becomes difficult to make comparisons between the studies and to assess the impact of reliability on validity. Only one empirical study directly examined the predictive validity of both consensus ratings and average ratings. Green et al. found only a very small difference between the predictive validity using consensus and panel averages.[150] For their 35-item measure, the predictive validity for consensus and panel average were .42 and .44 respectively. For the 8-item subset of critical tasks, the validity was .40 and .43. The authors concluded that validity was not affected by whether group consensus or average scores was used. They did, however, suggest that other methods for combining individual scores be investigated, as they believed the choice between the two should not be a matter of indifference.

Most remaining studies of structured panel interviews used consensus as the combination method for validity and reported a range of –.04[151] to as high as .81.[152] Only two studies employed statistical average and also investigated predictive validity. They obtained a predictive validity range of .26[153] to .44.[154]

Pulakos & Schmitt required that raters evaluate each interviewee independently upon completion of the interview.[155] Then, interviewers discussed the ratings with each other and came to a consensus regarding how the interviewees would be rated on each dimension. Interviewers were instructed not to change independent ratings as a result of consensus discussions. The initial independent ratings were retained for calculating reliability, while the consensus rating was utilized for validity measures. Latham, et al. employed the same procedure and methods for both reliability and predictive validity.[156]

Landy reported that at the end of the interview, each panel member independently rated the applicants on nine dimensions of a structured interview.[157] On the basis of these ratings, each panel member made a hiring recommendation. Both the averaged trait ratings and the individual dimensions were used as predictor variables. Dimensions seem to predict performance better than the overall recommendation. The overall recommendation yielded non-significant validity against all factors of performance, whereas dimensions of interview ratings showed some predictive validity ranging from .26 to .34.

In sum, only one study directly compared rating method with predictive validity.[158] Their results were inconclusive. Furthermore, only two studies used statistical average as the rating combination, making comparisons difficult and tenuous. Utilizing the available empirical and meta-analytic studies, it appears that combination method results are largely equivocal. Statistical averages seem to produce slightly higher reliabilities, but validity results are inconclusive. The relationship of that combination method has with interview reliability and validity must for technical accuracy be resolved through future research.

Step Eight: Evaluate selection decisions based on subsequent employee performance

Description and Explanation

Panel interview studies vary in the measures they utilize to assess performance. Some studies attempted to predict early training performance,[159] whereas others predicted later training exercise performance.[160] One study attempted to predict time to promotion.[161] Most studies, however, attempted to predict on-the-job performance within a range of 3 to 12 months after hiring. Interviewees were rated by peers, supervisors, or both.

A recent trend in performance criteria is the use of behavioral observation scales (BOS)[162] or behavioral summary scales.[163] These scales are developed in accordance with job analysis and critical incident techniques. In essence, critical incidents are derived from the job analysis. Then, these incidents are combined to create observable traits that can be rated by supervisors or peers. The advantage of such scales is that they are "based on overt employee behavior rather than traits or economic constructs."[164] The use of these scales appears to be positive in terms of comparability of studies and legal defensibility of the interview process.[165] The scales seem to be the current preferred method of performance appraisal in interview studies.

Review

One of the greatest difficulties in reviewing and comparing panel interview studies is a lack of consistency in performance criteria. The studies attempt to assess the predictive ability of panel interviews, but they lack a consensus as to what they are trying to predict. The importance of performance criteria is highlighted in several studies. For example, Gardner and Williams reported predictive validity of .34–.37 for early training performance, and .30 for specialist performance, but found negative predictive ability (–.25–.31) for time to promotion.[166] In contrast, Roth and Campion reported a validity of .20 for promotion and .41 for performance, but a non-significant validity for training.[167] Landy reported significant validities for individual dimensions, but failed to find significant predictive ability for overall performance.[168] Borman failed to find any significant predictive ability for performance on early or later exercises.[169] Thus, one can see that a difficulty in comparing studies is that each utilizes a different performance measure.

In terms of long-range predictability, Anstey conducted a 30-year longitudinal study of civil service personnel.[170] Based on a sample size of 301, he found a corrected predictive validity of .66. Using Vernon's[171] reported validity of .56 after two years of service, his own previous report of .60 after 17 years of service (conducted in 1964), and his current finding of .66 after 30 years of service, Anstey concluded that validities tend to increase as long-term criteria become available. However, Gardner & Williams' 25-year longitudinal study of British Royal Navy officers contradicts that finding.[172] They found that the two long-term criteria, namely, speed of promotion to commander and usefulness to the service, yielded validities of .15 and .14 respectively.

The were lower than validity for earlier performance measures. These differences may be explained by the different criteria used. However, additional, more strictly designed studies are needed to elucidate reasons for this discrepancy.

One positive trend is the use of behavioral observation scales (BOS) or behavioral summary scales.[173] Four studies strictly followed this procedure.174 Even though they did not use exactly the same scale, their studies provided validity information that can be compared easily and meaningfully. For example, most of the studies reported validities in the .30–.45 range. Latham and colleagues[175] reported significant validities for situational interviews and not behavior-based, while Pulakos and Schmitt reported the exact opposite.[176] Because of similar techniques, however, differences in meaningful terms (e.g., setting, training, etc.), can be explained instead of simply concluding they were measuring different outcomes.

While overall conclusions regarding the predictive ability of panel interviews are difficult to make in a narrative review, it is clear that performance scales derived from job analysis and based on observable, measurable traits are the preferred performance criteria, not only for legal defensibility,[177] but also for meaningful conclusions and comparisons between studies.

Conclusions and Directions for Future Research

Previous narrative and meta-analytic reviews have reported conflicting results concerning the panel interview. This review sought to explain and review features of the panel interview that may enhance its usefulness and psychometric soundness. An 8-step process of conducting the panel interview was utilized as a framework for reviewing the literature on panel interviews. This review revealed that setting, job analysis, scoring anchors, question type, training, structure, combination method, and predictive criteria have all demonstrated usefulness in explaining variance between panel interview studies. Based on the previous discussion of these features, the following conclusions and recommendations for future applied research are offered.

With regard to setting, research has demonstrated that panel interviews can be effectively utilized in either public or private settings. However, regardless of setting, almost all studies have been conducted at the entry level. It may be possible that panel interviews can better predict success of managerial level than entry-level positions (or vice versa), but virtually no study has investigated these higher levels in an organization. It is important to note, however, that the lack of studies dealing with managerial and executive positions is probably related more to sample size (more subjects are available for studies involving entry-level positions) than to the utility of panel interviews for these positions. Future research may want to address and compare panel interview outcomes at various organizational levels, as most current research has targeted only entry-level positions. In fact, the utility of a panel interview may be even greater when selecting people for higher-level positions. Since costs of poor selection decisions are generally greater at higher levels, it might be easier to justify the additional persons necessary for a panel format. Weston and Warmke argue that panel inter-

views may actually be more time efficient than the series of one-on-one interviews generally utilized when hiring for a managerial position.[178] Field testing of these factors is essential as laboratory experiments lend little insight into practical outcomes.

Structure and scoring anchors impact the reliability and validity of panel interviews. In general, structured panel interviews demonstrate higher reliability and validity than unstructured interviews. Scoring anchors are an important component of this structure, particularly in relation to reliability, and they seem to be increasingly utilized in both research and practice. Future studies need to directly compare the use of structured and unstructured panel interviews, holding constant the other variables such as scoring anchors and question type, and making sure both reliability and validity are noted for both types of interviews. Furthermore, useful future research might be to directly test the use of scoring anchors versus no scoring anchors. Lin et al. is a good example of this type of study, comparing a traditional structured interview with a situational one using scoring anchors.[179] However, this study has limited usefulness for comparison because it does not include a performance rating (for validity measures), nor does it compare independent ratings (for reliability). Still, a study similar to this, which specifically tests the use of scoring anchors, would be helpful to add information to the literature on interview reliability and validity. Finally, future studies also should determine more clearly the effects and utility of scoring anchors. Previous practical suggestions and this review suggest that 5-point scales may be superior to two points, but this difference has not been tested.[180] Laboratory experiments with different scoring scales would be a relatively simple method of testing this preliminary proposition.

Question type seems to have little correlation with inter-rater reliability, assuming that the questions are structured and perhaps based on specific scoring anchors. Although validity results seem conflicting, the controlled nature of the Campion et al.[181] and especially the Pulakos and Schmitt[182] studies lend compelling evidence to support the superiority of behavior-based panel interviews. Future research should extend studies like these into a field setting, comparing behavioral and situational interviews in real-life situations with actual consequences and performance ratings.

Meta-analysis has demonstrated training effects across individual and panel studies. Our review showed that training had some relationship to improved reliability, but did not appear to affect validity of panel interviews. One reason for this finding may be lack of information regarding the type of training and how it was conducted. We know little about the content or effectiveness of the training sessions. Future studies would add to the literature by expounding on the training their interviewers completed and also may directly compare trained and untrained panels.[183] Another possible direction in this area would be to investigate the benefits of training and utilizing "teams" of interviewers, which conduct large numbers of interviews as a group (as opposed to ad-hoc panels that change members for each job search). These teams, when trained together, may develop similar schemas or "mental models,"[184] which could enhance their ability to correctly and consistently assess candidates. Finally, the

panel interview literature would benefit from a moderator analysis similar to Huffcutt and Woehr with only panel studies.[185]

Combination method appears to be particularly problematic. Some studies use a consensus rating for reliability and validity, some use the statistical average for reliability and consensus for validity, and some use the average rating for both. This variation makes evaluation difficult. Even meta-analytic comparisons are suspect due to small sample size(s).[186] Future research may directly test these two combination methods. If statistical averages are utilized for assessing reliability, these scores simply could be retained for validity analysis. Then, the averaged scores and the consensus scores could be compared to assess which one produces the best predictions.

Make-up of the panel is related to the combination method. This is an interesting facet of panel interviews not often discussed in the literature that deserves more research attention. Lin et al. suggested that racial diversity of the panel might add to outcome fairness, especially when interviewing minority candidates.[187] Another aspect that seems critical is the amount of familiarity panel members have in doing interviews together. That is, the experience of the panel *as an ongoing panel* seems important and often overlooked. It would be interesting to investigate whether these teams tend to evolve toward "groupthink" or if they are likely to develop more critical discussion and decision-making skills that lead to more valid decisions. None of the studies reviewed in this paper note the amount of experience panelists had working together. Future research needs to draw on the team dynamics literature to explore possible experience effects.

Finally, the lack of consistency in predictive criteria utilized across the studies is problematic. Results are difficult to compare due to the wide range of criteria (and measures of it) used to assess performance. Future research would benefit from streamlining of these measures. Perhaps behavioral observation scales, reported by self, superior, and peers, may be obtained at set intervals (e.g., 2 weeks, 3 months, 6 months). This procedure would not limit the research and would aid in making comparisons between studies.

Practical Implications

Our analysis points to a number of practical implications for organizations interested in utilizing panel interviews. While this paper is certainly not the first to suggest these or other practical implications, we extend the previous suggestions in three critical ways.[188] First, the current paper provides a comprehensive review of the existing panel literature, so the scholar and practitioner can assess the available evidence when drawing conclusions. Second, although Pursell et al. proposed a highly structured approach that they supported in a later field study, the current review supports and extends many of their suggestions by providing additional empirical evidence and explaining differences between the studies.[189] Third, the current review provides additional information concerning setting, training, combination method, and performance criteria that have not been included in previous prescriptive literature.

First of all, research has demonstrated that panel interviews may be effectively utilized in both public and private settings. While primarily studied in entry-level positions, the panel interview is certainly applicable for all levels of the organization. Once the panel interview is chosen, the practitioner must conduct a thorough job analysis, then develop questions that directly relate to the critical components of the job.[190] While the recommendation for conducting a job analysis does not flow directly from this review, we would be remiss not to include it in a process summary because of its centrality to the content and face validity of job interviews. Situational questions initially demonstrated superiority to behavioral questions in terms of predicting job performance.[191] Later evidence, however, suggests that behavioral questions may be superior.[192] Both are relatively equal in terms of reliability. The current evidence points to behavioral questions as the preferred use.

Next, scoring anchors should be developed for at least high, medium, and low level responses.[193] There is some evidence that a 5-point scale (1=low, 3=medium, and 5=high) demonstrates the highest reliability. In addition, the scoring anchors should include sample responses and also keywords that cue the interviewers as to the level of response.[194] The responses aids in validity, by helping interviewers tune in to only job relevant information.

Then, the interview panel should be selected and trained in how to conduct a structured interview and how to utilize the scoring anchors. The number of interviewers has not been directly tested in the literature, but most prescriptive summaries recommend three to six.[195] Warmke and Weston recommend odd numbers so that the likelihood of split decisions is reduced.[196] Regarding training, the empirical research has provided little insight on how to conduct it, yet at least one study suggested that interviewers who perform poorly in training should be removed from the panel.[197] In addition, training provides a crucial pilot test of the instrument. If panels are consistently low in reliability and/or agreement during the training, this may alert question designers that there are problems with the questions or scoring anchors. Anchors that lack clarity have been shown to produce inconsistent results.[198]

While research has shown advantages in terms of reliability of utilizing statistical combination as opposed to consensus ratings, the validity of these different methods is unclear. Campion et al. suggest that rating each answer then combining them at the end is the highest level of structure.[199] Furthermore, Pursell et al. advise the use of cut scores before making a final decision. [200] They argue that simply ranking the candidate leads to the most adverse impact and leaves little room for affirmative action goals. Another approach that seems intuitively appealing might be to combine the methods. That is, instead of using only average ratings, or only a consensus rating as a predictor, panels could make use of a discussion period immediately following the interview, followed by individual, private ratings that would be combined into an overall average composite. Similar methods of statistical rating combination have been reviewed by assessment center researchers. [201] This method of combination would alleviate some drawbacks of each approach. For example, the panel could gain the benefits of discussion among members (e.g. pointing out attributes or contributions that other inter-

viewers might have overlooked, addressing biases, etc.), yet avoid pitfalls of consensus ratings (e.g., social facilitation or pressure to conform to a senior panel member). [202] Likewise, taking average individual ratings following the discussion period still allows each person to have an equal vote, along with anonymity in their ratings.

In sum, we recommend first establishing cut scores before the interview process. Then, after each interview, have each panel member record his/her rating of each candidate. Statistically combine these ratings to assess reliability of the instrument and, later, predictive validity. After the initial ratings are recorded, and all interviews should have been conducted, have the group discuss the candidates together. Finally, the interviewers should independently re-rate the individual, combine their scores into an overall composite, and compare the results to the cut score. Candidates above the cut score are retained for further consideration. Both the initial average and the overall consensus score should be retained for assessing validity of the instrument.

Performance evaluations should be based on the same analysis as the interview questions. Latham et al. suggest the use of behavioral observation scales because they are based on observable behaviors, not intangible traits. [203] While the best timing of performance appraisals is not clear, initial ratings at three months, six months, and one year seem to provide a reasonable timeline for tracking progress.

In addition to the importance of organizational outcomes predicated on quality selection decisions, firms also must be aware of legal and moral implications surrounding these assessments. Regarding fairness and minority characteristics, Reilly and Chao found no evidence that interviews have less adverse impact than tests. [204] In fact, Pursell et al. suggested that panel interviews were potentially one way of reducing bias in the selection process. [205] In our review, Latham et al. reported no discrimination against blacks and women in their study of sawmill applicants. [206] Lin et al. showed stronger same-race effects with the conventional structured interviews than with the situational interview. [207] It was suggested that both adding structure to the interview and using mixed race interview panels could minimize same-race bias. No age similarity effects were detected with either procedure. While more research in the area of minority characteristics in panel interviews (especially the issue of adverse impact) is certainly called-for, current research indicates that panel interviews may be promising as a fair and legally defensible selection tool. [208]

Although it has been utilized for over 50 years, the panel interview remains somewhat controversial as a valid selection tool. The current review attempts to illuminate reasons behind the differences of opinion. Because of the potential of panel interviews for reduced bias and the addition of a more social side to selection decisions, they must continue to be investigated and improved as an important selection tool.

Notes

[1] Pursell, E. D., Campion, M. A., & Gaylord, S. R. 1980, November. Structured interviewing: Avoiding selection problems. *Personnel Journal*, 907–912; Weston, D. J., & Warmke, D. L. 1988. Dispelling myths about panel interviews. *Personnel Administrator*, 5: 109–111.

[2] Pursell et al., op cit.; Warmke, D. L., & Weston, D. J. 1992. Success dispels myths about panel interviews. *Personnel Journal*, 4: 120–130; Weston & Warmke, op cit.

[3] Weston & Warmke, op cit.

[4] Campion, M. A., Palmer, D. K., & Campion, J. E. 1997. A review of structure in the selection interview. *Personnel Psychology*, 50: 655–702; Jelf, G. S. 1999. A narrative review of post–1989 employment interview research. *Journal of Business and Psychology*, 14: (1) 25–58; Huffcutt, A. I., & Woehr, D. J. 1999. Further analysis of employment interview validity: A quantitative evaluation of interviewer-related structuring methods. *Journal of Organizational Behavior*, 20: 549–560; Marchese, M. C., & Muchinsky, P. M. 1993. The validity of the employment interview: A meta-analysis. *International Journal of Selection and Assessment*, 1: 18–26; McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. 1994. The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, 79: 599–616; Wiesner, W. H., & Cronshaw, S. F. 1988. A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology*, 61: 275–290.

[5] Wiesner & Cronshaw, op cit.

[6] Huffcutt & Woehr, op cit.

[7] Marchese & Muchinsky, op cit.

[8] For examples see Mayfield, E. C. 1964 The selection interview: A reevaluation of published research. *Personnel Psychology*, *17*: 239–260; Campion et al., 1997, op cit.

[9] For a review of a number of selection devices see Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35*, 1–62.

[10] Arvey, R. D., & Campion, J. E. 1982. The employment interview: A summary and review of recent research. *Personnel Psychology, 35*: 281–318; Campion, M. A., Pursell, E. D., & Brown, B. K. 1988. Structured interviewing: Raising the psychometric properties of the employment interview. *Personnel Psychology, 41*: 25–42; Pursell et al., op cit.; Warmke & Weston, op cit.; Weston & Warmke, op cit.

[11] Mayfield, op cit.; Weston & Warmke, op cit.

[12] Glaser, R., Schwarz, P. A., & Flanagan, J. C. 1958. The contribution of interview and situational performance procedures to the selection of supervisory personnel. *Journal of Applied Psychology, 42*: 69–73.

[13] For examples see Campion, M. A., Campion, J. E., & Hudson, J. P. 1994. Structured interviewing: A note on incremental validity and alternative question types. *Journal of Applied Psychology, 79*: 998–1002; Pulakos, E. D., & Schmitt, N. 1995. Experience-based and situational interview questions: Studies of validity. *Personnel Psychology, 48*: 289–308.

[14] Campion et al., 1997, op cit., p. 681.

[15] For police settings see Davey, B.1984. Are all oral panels created equal? A study of differential validity across oral panels. Paper presented at the IPMA Assessment Council Annual Conference; DuBois, P. H., & Watson, R. I. 1949. The selection of patrolmen. *Journal of Applied Psychology, 34*: 90–95; Flynn, J. T., & Peterson, M. 1972. The use of regression analysis in police patrolman selection. *The Journal of Criminal Law, 63*: 564–569; Green, P. C., Alter, P., & Carr, A. F. 1993. Development of standard anchors for scoring generic past-behavior questions in structured interviews. *International Journal of Selection and Assessment, 1*: 203–212; Landy, F. J. 1976. The validity of the interview in police officer selection. *Journal of Applied Psychology, 61*: 193–198; Reynolds, A. H. 1979. The reliability of a scored oral interview for police officers. *Public Personnel Management, 8*: 324–328. For military settings see Borman, W. C. 1992. Validity of behavioral assessment for predicting military recruiter performance. *Journal of Applied Psychology, 67*: 3–9. Gardner, K. E., & Williams, A. P. O. 1973. A twenty-five year follow-up of an extended interview selection procedure in the Royal Navy. *Occupational Psychology, 47*: 1–13; Glaser et al., op cit.

[16] Anstey, E. 1977. A 30-year follow-up of the CSSB procedure, with lessons for the future. *Journal of Occupational Psychology, 50*: 149–159. Pulakos & Schmitt, op cit.; Stohr-Gillmore, M. K., Stohr-Gillmore, M. W., & Kistler, N. 1990. Improving selection outcomes with the use of situational interview: Empirical evidence from a study of correctional officers of new generation jails. *Review of Public Personnel Administration, 10*: 1–18. Vernon, E. P. 1950. The validation of Civil Service Selection Board procedures. *Occupational Psychology, 24*: 75–95.

[17] Dubois & Watson, op cit.

[18] Anstey, op cit.; Dubois & Watson, op cit.; Flynn & Peterson, op cit.; Gardener & Williams, op cit.; Glaser et al., op cit.; Landy, op cit.; Reynolds, op cit.; Vernon, op cit.

[19] Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. 1980. The situational interview. *Journal of Applied Psychology, 65*: 422–427.

[20] Ibid.

[21] For public settings see Borman, op cit.; Green et al., op cit.; Pulakos & Schmitt, op cit.; Stohr-Gillmore et al., op cit. For private settings see Campion et al., 1988, op cit.; Latham et al., op cit.; Latham, G. P., & Saari, L. M. 1984. Do people do what they say? Further studies on the situational interview. *Journal of Applied Psychology, 69*: 569–573; Lin, T. R., Dobbins, G. H., & Farh, J. L. 1992. A field study of race and age similarity effects on interview ratings in conventional and situational interviews. *Journal of Applied Psychology, 77*: 363–371; Roth, P. L., & Campion, J. E. 1992. An analysis of the predictive power of the panel interview and pre-employment tests. *Journal of Occupational and Organizational Psychology, 65*: 51–60.

[22] Warmke & Weston, op cit.; Weston & Warmke, op cit.

[23] Huffcutt & Woehr, op cit.; Marchese & Muchinsky, op cit.

[24] Huffcutt & Woehr, op cit.

[25] Wiesner & Cronshaw, op cit.

[26] McDaniel et al., op cit.

[27] Huffcutt & Woehr, op cit.

[28] Wiesner & Cronshaw, op cit.

[29] Marchese & Muchinsky, op cit., p. 29.

[30] Huffcutt & Woehr, op cit.; McDaniel et al., op cit.; Wiesner & Cronshaw, op cit.

[31] Wiesner & Cronshaw, op cit.

[32] Huffcutt & Woehr, op cit.

[33] Campion et al., 1988, op cit; Pursell et al., op cit.; Warmke & Weston, op cit.

[34] Pursell et al., op cit.

[35] For examples see Dubois & Watson, op cit.; Glaser et al., op cit.

[36] Campion et al., 1994, op cit.; Campion et al., 1997; op cit.; Latham et al., op cit.; Latham & Saari, op cit.; Pulakos & Schmitt, op cit.

[37] Campion et al., 1994; op cit.

[38] Campion et al., 1994; op cit.; Campion et al., 1997; op cit.; Pulakos & Schmitt, op cit.

[39] Campion et al., 1994, op cit.; Campion et al., 1997; op cit.; Latham et al., op cit.; Latham & Saari, op cit.; Pulakos & Schmitt, op cit.; Stohr-Gilmore et al., op cit.

[40] Campion et al., 1997; op cit.

[41] Latham et al., op cit.; Latham & Saari, op cit.

[42] Campion et al., 1994, op cit.

[43] Pulakos & Schmitt, op cit.

[44] Campion et al., 1997; op cit.

[45] Campion et al., 1997; op cit.; Pulakos & Schmitt, op cit

[46] Latham & Saari, op cit.

[47] Latham et al., op cit.

[48] Pulakos & Schmitt, op cit.

[49] Latham & Saari, op cit.

[50] Lin et al., op cit.

[51] Green et al., op cit.

[52] Stohr-Gillmore et al., op cit.

[53] Landy, op cit.

[54] Glaser et al., op cit.

[55] Green et al., op cit.

[56] Pulakos & Schmitt, op cit.

[57] Latham & Saari, op cit.

[58] Latham et al., op cit.

[59] Ibid.

[60] Campion et al., 1994, op cit.

[61] Pulakos & Schmitt, op cit.

[62] Campion et al, 1994 op cit.

[63] In the Campion et al., 1994 study, the interviewees were asked both types of questions.

[64] Pulakos & Schmitt, op cit.

[65] Campion et al., 1994, op cit.

[66] Campion et al., 1997 op cit.; Huffcutt & Woehr, op cit.; Latham & Saari, op cit.; Latham et al., op cit.; Pulakos & Schmitt, op cit.

[67] Campion et al., 1988, op cit.; Campion et al., 1997 op cit.; Green et al., op cit.; Landy, op cit.; Latham et al., op cit.; Pulakos & Schmitt, op cit.; Reynolds, op cit.; Roth & Campion, op cit.; Stohr-Gillmore et al., op cit.

[68] Campion et al., 1997 op cit.; Green et al., op cit.; Latham et al., op cit.; Pulakos & Schmitt, op cit.; Roth & Campion, op cit.; Stohr-Gillmore et al., op cit.

[69] Campion et al., 1997 op cit.; Latham & Saari, op cit.; Latham et al., op cit.; Pulakos & Schmitt, op cit.

[70] Davey, op cit.

[71] Lin et al., op cit.

[72] Green et al., op cit.

[73] Stohr-Gillmore et al., op cit.

[74] Pulakos & Schmitt, op cit.

[75] Latham & Saari, op cit.; Latham et al., op cit.

[76] Latham & Saari, op cit.

[77] Campion et al., 1988, op cit.

[78] Green et al., op cit.

[79] Latham & Saari, op cit.

[80] Landy, op cit.

[81] Latham & Saari, op cit.; Latham et al., op cit.

[82] Pulakos & Schmitt, op cit.

[83] Marchese & Muchinsky, op cit.

[84] Campion et al., 1997, op cit.

[85] Gardner & Williams, op cit.

[86] Warmke & Weston, op cit.

[87] Huffcutt & Woehr, op cit., p 549.

[88] Davey, op cit.

[89] Lin et al., op cit.

[90] Roth & Campion, op cit.

[91] Green et al., op cit.

[92] Pulakos & Schmitt, op cit.

[93] Davey, op cit.; Lin et al., op cit.

[94] Green et al., op cit.; Pulakos & Schmitt, op cit.

[95] Davey, op cit.

[96] Lin et al., op cit.; Latham & Saari, op cit.

[97] Davey, op cit.

[98] Lin et al., op cit.

[99] Davey, op cit.; Green et al., op cit.; Pulakos & Schmitt, op cit.

[100] Roth & Campion, op cit.

[101] Green et al., op cit,

[102] Pulakos & Schmitt, op cit.

[103] Lin et al., op cit.

[104] Davey, op cit.

[105] Green et al., op cit.

[106] Pulakos & Schmitt, op cit.

[107] Ibid.

[108] Davey, op cit.; Pulakos & Schmitt, op cit.

[109] Davey, op cit.

[110] Roth & Campion, op cit.

[111] Huffcutt & Woehr, op cit.

[112] Campion et al., 1988, op cit.

[113] Campion et al., 1997, op cit.; Latham & Saari, op cit.; Latham et al., op cit.; Pulakos & Schmitt, op cit.

[114] Campion et al., 1988, op cit.

[115] Campion et al., 1997, op cit.; Huffcutt & Woehr, op cit.; Marchese & Muchinsky, op cit.; Wiesner & Cronshaw, op cit.

[116] Campion et al., 1997; op cit.

[117] Campion et al., 1997, op cit.; Marchese & Muchinsky, op cit.; Wiesner & Cronshaw, op cit.

[118] Wiesner & Cronshaw, op cit.

[119] Stohr-Gillmore et al., op cit.

[120] Lin et al., op cit.

[121] Glaser et al., op cit.

[122] Lin et al., op cit.

[123] Davey, op cit.; Lin et al., op cit.

[124] Glaser et al., op cit.

[125] Campion et al., 1997, op cit.

[126] Campion et al., 1988, op cit.; Landy, op cit.; Latham & Saari, op cit.; McDaniel et al., op cit.; Wiesner & Cronshaw, op cit.

[127] Pulakos & Schmitt, op cit.

[128] Green et al., op cit.

[129] Latham & Saari, op cit.

[130] Latham et al., op cit.

[131] Glaser et al., op cit.

[132] Campion et al., 1988, op cit. reported a value of .56 and Green et al., op cit. reported a value of .81.

[133] Wiesner & Cronshaw, op cit.

[134] McDaniel et al., op cit.

[135] Pulakos & Schmitt, op cit.

[136] Latham & Saari, op cit.; Latham et al., op cit.

[137] Stohr-Gillmore et al., op cit.

[138] Campion et al., 1997, op cit.

[139] Pulakos & Schmitt, op cit.

[140] Latham & Saari, op cit.; Latham et al., op cit.

[141] Janis, I.L. 1982. *Victims of Groupthink (2nd ed.)*. Boston: Houghton-Mifflin.

[142] Geen, R.G. 1991. Social motivation. *Annual Review of Psychology, 42*: 377–399; Seta, J.J., Wang,. M.A., Crisson, J.E., & Seta, C.E. 1989. Audience composition and felt anxiety: Impact averaging and summation. *Basic Applied Social Psychology, 10*: 57–72.

[143] Wiesner & Cronshaw, op cit.

[144] Reynolds, op cit.

[145] Landy, op cit.

[146] Excluding those studies that used post-discussion inter-rater reliability, i.e. Davey, op cit. and Lin et al., op cit.

[147] Stohr-Gillmore et al., op cit.

[148] Latham & Saari, op cit.

[149] Davey, op cit. and Lin et al., op cit.

[150] Green et al., op cit.

[151] Stohr-Gillmore et al., op cit.

[152] Green et al., op cit.

[153] Landy, op cit.

[154] Green et al., op cit.

[155] Pulakos & Schmitt, op cit.

[156] Latham et al., op cit.

[157] Landy, op cit.

[158] Green et al., op cit.

[159] Gardner & Williams, op cit.

[160] Borman, op cit.

[161] Gardner & Williams, op cit.

[162] Latham, G.P., & Wexley, K. N. 1977. Behavioral observation scales for performance appraisal purposes. *Personnel Psychology, 30*: 255–268.

[163] Pulakos & Schmitt, op cit.

[164] Latham et al., op cit., p. 426.

[165] Campion et al., 1988, op cit.

[166] Gardner & Williams, op cit.

[167] Roth & Campion, op cit.

[168] Landy, op cit.

[169] Borman, op cit.

[170] Anstey, op cit.

[171] Vernon, op cit.

[172] Gardner & Williams, op cit.

[173] Latham & Wexley, op cit.; Pulakos & Schmitt, op cit.

[174] Campion et al., 1988, op cit.; Latham & Saari, op cit.; Latham et al., op cit.; Pulakos & Schmitt, op cit.

[175] Latham et al., op cit.

[176] Pulakos & Schmitt, op cit.

[177] Campion et al., 1988, op cit.

[178] Weston & Warmke, op cit.

[179] Lin et al., op cit.

[180] For example, see Campion et al., 1988, op cit.

[181] Campion et al., 1994, op cit.

[182] Pulakos & Schmitt, op cit.

[183] Green et al., op cit.

[184] Klimoski, R. & Mohammed, S. 1994. Team mental model: construct of metaphor? *Journal of Management 20*: 403–437.

[185] Huffcutt & Woehr, op cit.

[186] Wiesner & Cronshaw, op cit.

[187] Lin et al., op cit.

[188] For examples see Campion et al., 1988, op cit.; Pursell et al., op cit.; Warmke & Weston, op cit.; Weston & Warmke, op cit.

[189] Campion et al., 1988, op cit.; Pursell et al., op cit.

[190] Latham et al., op cit.; Pursell et al., op cit.

[191] Latham & Saari, op cit.; Latham et al., op cit.

[192] Campion et al., 1994, op cit.; Pulakos & Schmitt, op cit.

[193] Pursell et al., op cit.

[194] Pulakos & Schmitt, op cit.

[195] Campion et al., 1988, op cit.; Warmke & Weston, op cit.

[196] Warmke & Weston, op cit.

[197] Davey, op cit.

[198] Latham et al., op cit.

[199] Campion et al., 1997, op cit.

[200] Pursell et al., op cit.

[201] For example Thornton, G.C. 1992. *Assessment Centers in Human Resource Management*. Reading, MA: Addison-Wesley Publishing Co.

[202] Geen et al., op cit.; Seta et al., op cit.

[203] Latham et al., op cit.

[204] Reilly & Chao, op cit.

[205] Pursell et al., op cit.

[206] Latham et al., op cit.

[207] Lin et al., op cit.

[208] Pursell et al., op cit.

Authors

Marlene Dixon
College of Education
The Ohio State University
Columbus, OH 43210

Sheng Wang
College of Business
The Ohio State University
Columbus, OH 43210

Jennifer Calvin
College of Education
The Ohio State University
Columbus, OH 43210

Brian Dineen
College of Business
The Ohio State University
Columbus, OH 43210

Edward Tomlinson
College of Business
The Ohio State University
Columbus, OH 43210