1997

Effect of applicant order on reliability and evaluations in the panel interview

Susan M. Howard
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_theses

Recommended Citation

Howard, Susan M., "Effect of applicant order on reliability and evaluations in the panel interview" (1997). *Master's Theses*. 1450.

DOI: <https://doi.org/10.31979/etd.syeh-r34m>

https://scholarworks.sjsu.edu/etd_theses/1450

This Thesis is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Theses by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

EFFECT OF APPLICANT ORDER ON RELIABILITY
AND EVALUATIONS IN THE PANEL INTERVIEW

A Thesis

Presented to

the Faculty of the Department of Psychology

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Susan M. Howard

May, 1997

UMI Number: 1384699

UMI Microform 1384699
Copyright 1997, by UMI Company. All rights reserved.

**This microform edition is protected against unauthorized
copying under Title 17, United States Code.**

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

APPROVED FOR THE DEPARTMENT OF PSYCHOLOGY

Howard Tokunaga, Ph.D.

Rebecca Malatesta, Ph.D.

Julie Kwan, Ph.D., Kaiser Permanente Medical Care Program

APPROVED FOR THE UNIVERSITY

© 1997

Susan M. Howard

ALL RIGHTS RESERVED

ABSTRACT

EFFECT OF APPLICANT ORDER ON RELIABILITY
AND EVALUATIONS IN THE PANEL INTERVIEW

by Susan M. Howard

The purpose of this study was to explore how the relative position an applicant occupied in a slate of sequential panel interviews would affect interrater reliability, the value of the interview ratings the applicant received, and the likelihood that the applicant would obtain a favorable interview outcome. Applicants were interviewed by 2 - 4 interviewers in 1 of 5 time slots, rated on leadership dimensions, and given a final recommendation. Data from 155 interviews were analyzed.

Results showed that interrater reliability of the final recommendation was higher for the last slot when compared to the first slot. The applicant's position in the interview slate did not significantly affect the value of average scores. However, applicants were significantly less likely to receive a positive final recommendation as the interview day progressed. These findings have implications for conducting panel interviews in ways that will ensure the fairest possible treatment for applicants.

TABLE OF CONTENTS

SECTION	PAGE
INTRODUCTION.....	3
Panel Interviews	5
Validity	5
Interrater reliability.....	6
Method of producing final interview outcomes.....	7
Applicant order.....	9
Purpose of This Study	12
METHOD	15
Overview	15
Applicants.....	17
Interviewers	17
Interview Design	18
Dimensions	18
Questions.....	18
Rating forms.....	19
Interview Administration.....	19
Overview	19
Interviewer preparation	20
Interview.....	21
Evaluation.....	21
Final decisions	21

RESULTS	22
Dependent Variables.....	22
Interrater Reliability (Hypothesis 1).....	22
Mechanically Combined Ratings (Hypothesis 2)	24
Subjective Recommendations (Hypothesis 3).....	27
DISCUSSION	31
Explanation of Results.....	32
Practical Implications.....	36
Strengths and Limitations	37
Future Research	38
Conclusion.....	41
REFERENCES	42
APPENDIXES	45
Appendix A. Rating Form.....	45
Appendix B. Human Subjects Approval Letter.....	51

LIST OF TABLES

TABLE	PAGE
1. Interrater Reliability of Average Score.....	23
2. Interrater Reliability of Recommendation.....	25
3. Average Score for Each Interview Slot	25
4. Number and Percentage of Recommendations for Each Interview Slot.....	28
5. Number and Percentage of Recommendations for Each Interview Slot Group	30

LIST OF FIGURES

FIGURE	PAGE
1. Average score for each interview slot.....	26
2. Recommendations for each interview slot.....	29

Running head: EFFECT OF APPLICANT ORDER IN THE PANEL INTERVIEW

Effect of Applicant Order on Reliability and Evaluations in the Panel Interview

Susan M. Howard

San Jose State University

Author Note

Correspondence concerning this study should be addressed to Susan M. Howard,
Department of Psychology, San Jose State University, San Jose, California 95192.

Abstract

The purpose of this study was to explore how the relative position an applicant occupied in a slate of sequential panel interviews would affect interrater reliability, the value of the interview ratings the applicant received, and the likelihood that the applicant would obtain a favorable interview outcome. Applicants were interviewed by 2 - 4 interviewers in 1 of 5 time slots, rated on leadership dimensions, and given a final recommendation. Data from 155 interviews were analyzed. Results showed that interrater reliability of the final recommendation was higher for the last slot when compared to the first slot. The applicant's position in the interview slate did not significantly affect the value of average scores. However, applicants were significantly less likely to receive a positive final recommendation as the interview day progressed. These findings have implications for conducting panel interviews in ways that will ensure the fairest possible treatment for applicants.

Effect of Applicant Order on Reliability and Evaluations in the Panel Interview

Today's hiring managers can choose from a number of devices in their selection arsenal, including biodata analysis, work samples, reference checks, assessment centers, drug tests, and formal inventories scoring everything from honesty to knowledge of insurance law to personality. Despite the plethora of choices available and concerns raised by personnel psychologists regarding the usefulness of the interview as a selection device (Arvey & Campion, 1982; Hunter & Hunter, 1984; Mayfield, 1964; Reilly & Chao, 1982, Schmitt, 1976; Ulrich & Trumbo, 1965; Wagner, 1949), the interview continues to be one of the most widely used selection methods. A survey of 245 organizations revealed that most applicants are interviewed at least twice in the process of being hired, regardless of level in the organization (Bureau of National Affairs, Inc., 1988). According to 444 personnel administrators who were surveyed about their selection practices, the outcome of the interview contributes the most to the selection decision, even if resumes, reference checks, or tests also are utilized (Cox, Schlueter, Moore, & Sullivan, 1989).

Managers have always liked and used interviews. One reason is that interviews can be used for multiple purposes besides evaluating the applicant (Arvey & Campion, 1982; Dipboye, as interviewed by Church, 1996); for example, companies can use interviews as a recruitment tool or public relations vehicle. Another reason is that managers believe themselves to be good judges of human character, and so want to have the chance to evaluate the applicant themselves, face-to-face (Arvey & Campion, 1982; Ferris, as interviewed by Church, 1996; Pulakos, as interviewed by Church, 1996; Roth, as interviewed by Church, 1996). Another attractive aspect of the interview for managers is that applicants expect to be assessed by an interview, and thus are less likely to challenge its job relevance in a lawsuit. Interviews are also generally easier for companies to administer

than other devices such as paper-and-pencil tests (Arvey & Campion, 1982) because a traditional employment interview can be conducted with little advance preparation. Finally, managers may be aware that in more recent years research has redeemed the interview as a viable selection device, provided it is structured or standardized (Campion, Pursell, & Brown, 1988; Huffcutt & Arthur, 1994; Wiesner & Cronshaw, 1988; Wright, Lichtenfels, & Pursell, 1989).

One form of the interview that is gaining in popularity is the *panel interview*, which involves two or more interviewers assessing an applicant in the same interview. Bell (1989) reported that in 1987, interviews with two interviewers “may account for as many as 40% of all first-contact job interviews” (p. 89), and an executive recruiter estimated that 10 - 20% of her clients encountered panel interviews of three or more interviewers (Bamford, 1986). Panel interviews are probably attractive to managers because (a) panel interviews have been recommended by personnel psychologists as part of the structured interview process (Campion et al., 1988); (b) managers believe that using panel interviews will increase the validity of the interview, because they increase involvement of key individuals in the hiring process (Warmke & Weston, 1992); and (c) panel interviews can be more economical than multiple single interviews (Weston & Warmke, 1988).

Much of the interview research to date has focused on individual interviews, whereas the unique aspects of the panel interview have not been fully explored. For example, panel interviews often take the form of a single group of people evaluating multiple applicants sequentially over the course of a day. Conducting interviews in this manner may create conditions that affect how useful panel interviews can be. Research that examines how interviewers’ ability to make decisions is affected by conducting these multiple panel interviews could reveal ways to maximize the utility of these types of

interviews for selection.

The purpose of this study is to explore how the relative position an applicant occupies in a slate of sequential panel interviews might affect reliability of ratings among interviewers, the value of the interview ratings the applicant receives, and the likelihood that the applicant will obtain a favorable interview outcome.

Panel Interviews

Having more than one interviewer conduct an interview at the same time is a practical alternative to interviewing applicants in a series of multiple, individual interviews because applicants do not have to spend an entire day interviewing, yet the appropriate decision-makers can all have a chance to evaluate the applicant and make a decision following the interview. The panel interview also has been recommended as an important step in structuring the interview and increasing its validity (Campion et al., 1988).

Validity. There are several means by which the panel interview is supposed to bring about better validity than individual interviews. Because panel interviews aggregate the interviewers' individual opinions, the biases of each of the single interviewers should be balanced or canceled out (Campion et al., 1988; Dipboye, 1992). Also, having multiple interviewers increases the number of unique perspectives available to the group, resulting in a better quality decision (Dipboye, 1992; Weston & Warmke, 1988). The group discussion that is often a part of panel interviews may serve to focus interviewers on the more job-relevant factors of the interview, thus improving validity (Gatewood, 1990). Finally, interviewers have all observed the same behavior and so have common foundations on which to base their ratings and discussion of the applicant (Gatewood, 1990). This is not possible with multiple individual interviews, even if they are structured.

The issue of whether panels increase the validity of the interview has been examined empirically, and so far the research indicates that panel interviews do not fulfill their promise. In a systematic meta-analysis comparing individual interviews to *board interviews* (i.e., panel interviews), Wiesner and Cronshaw (1988) found that there was no difference in validity between these two formats when collapsed across degree of structure. However, panel interviews were more valid than individual interviews when they were unstructured. In another meta-analysis, McDaniel, Whetzel, Schmidt, and Maurer (1994) found that panel interviews are actually less valid than individual interviews when matched according to their degree of structure. In determining validity, it may be that the unique dynamics of panel interviews mentioned above are not as important as the structure built into them.

In spite of the mixed reviews of the panel interview's validity, it is still worth researching. One reason is that other than structure, possible moderators of panel interview validity have not been taken into account by the meta-analyses discussed here. For example, factors such as interviewer backgrounds (some suggestions for optimal panel construction are provided in Campion et al., 1988 and Warmke & Weston, 1992) and whether the panel as a whole had a chance to discuss the applicant may also affect the validity of panel interviews. Secondly, as stated earlier, panels are often used in organizations and appear to be gaining in popularity (Bamford, 1986; Bell, 1989). Given the pervasiveness of panel interviews in corporate America, researchers ought to find ways of improving their usefulness.

Interrater reliability. Research also has been conducted on interrater reliability in the panel interview. Reliability is another important measure of the quality of a selection tool because it imposes a limit on how valid the tool can be (i.e., an unreliable instrument

will not achieve respectable validity). To arrive at a measure of interrater reliability in a panel interview, the ratings an applicant receives from different interviewers in the same interview are compared. As discussed by Conway, Jako, and Goodman (1995), this panel interview coefficient is unique because it is a pure reflection of “disagreement over interpretation and evaluation of applicant performance” (p. 566), whereas a separate interview coefficient reflects both interviewer disagreement and real differences in applicant performance in the separate interviews. In fact, this is why Conway et al. hypothesized (and discovered) higher reliabilities for panel interviews (.77) than for separate interviews (.53) in their meta-analysis.

Standardization and structure have been investigated as possible moderators of interrater reliability in the panel interview. Conway et al. (1995) found that standardizing interview questions positively affected the interrater reliability of panel interviews, perhaps because the questions were able to elicit a more job-related sample of applicant behavior in the interview. In an early study where panel interviewers were asked to rank applicants, Schwab and Heneman (1969) reported that the reliability of the rankings was .79 for a structured interview, .43 for a semistructured interview, and .36 for an unstructured interview. In support of these findings, a meta-analysis which included 186 interrater reliability coefficients showed that structure moderates the reliability of panel interviews (McDaniel et al., 1994). Structured panel interviews had a mean reliability of .84, whereas unstructured panel interviews had a mean reliability of .68. Aside from these few studies, there does not appear to be other research on moderators of interrater reliability in the panel interview.

Method of producing final interview outcomes. There are many possible methods to use for producing final outcomes from a panel interview. The three main dimensions

along which methods vary are (a) how many evaluations are generated for each applicant (for example, a single global determination, ratings on several dimensions, or ratings on each question), (b) whether the final outcomes are to consist of mechanically combined ratings (for example, ratings are summed or inserted into a weighted formula) or clinical (i.e., subjective) decisions by the interviewers, and (c) whether interviewers are to be allowed to discuss ratings prior to producing final outcomes. For example, interviewers can rate the applicant individually on multiple dimensions, and then ratings can be mechanically combined to arrive at a final rating of the applicant. Alternatively, interviewers can discuss their ratings on dimensions and settle on final group ratings on the same dimensions. In yet another scenario, interviewers can be asked to discuss ratings before giving the applicant a single positive or negative recommendation.

In their study of interview reliability, Conway et al. (1995) found that the more evaluations are generated for each applicant, the higher the reliability (addresses dimension *a* above). Also, mechanically combined ratings were clearly superior to subjective ratings in terms of reliability (addresses dimension *b* above), especially when the number of ratings to be integrated before coming to a decision increased. This last finding supports Dawes (1979), who asserts that statistical (mechanical) combination of criteria results in higher quality decisions than subjective judgments of the same criteria, even when using expert raters.

Wiesner and Cronshaw (1988) found that consensus ratings were more valid than statistically combined individual ratings for the panel interview (addresses dimension *c* above). According to Pulakos, Schmitt, Whitney, and Smith (1996), very little research has been done on consensus versus average ratings. When exploring this issue in their study, they also discovered that consensus ratings are significantly more valid than

averaged individual ratings, but conclude that the difference is small and may not justify the additional time the interviewers must spend to reach consensus.

Applicant order. In their review of the interviewing literature, Arvey and Campion (1982) presented a three-pronged model of the factors and processes at work in the employment interview. The two areas that have received the most research attention have been applicant factors and interviewer factors, whereas the third area, situational factors (e.g., degree of interview structure, physical setting of the interview), has not been as well researched. For multiple sequential panel interviews, the order in which interviewers evaluate applicants is an important situational factor.

One aspect of applicant order that has been researched heavily is how interviewers' judgment of an applicant for a position might be affected by their prior judgment of another applicant. These so-called *context effects* can either be in the direction of prior applicants (an *assimilation effect*) or in opposition to prior applicants (a *contrast effect*). Of the two effects, contrast appears to be more in evidence (Dipboye, 1992). Research on contrast effects in the interview has been under investigation since the late 1960s, although not as much research has been done on the topic recently.

Contrast effects typically have been elicited by asking subjects to judge multiple job applicants sequentially. The order and quality of the applicants have been manipulated by the experimenters to produce in the subject's mind a "contrast" between the qualifications of the applicant presently under consideration and those seen before. By systematically varying the order of applicants who differ in quality, researchers can examine whether applicants are rated higher or lower depending on their position in the applicant line-up.

Most researchers have found that decision-makers' ratings do reflect a tendency to contrast the current applicant with the ones who came before, over and above any real

differences in applicant quality. Wexley, Yukl, Kovacs, and Sanders (1972) found that when an average-quality applicant was seen after two high- or low-quality applicants, 80% of the variance in subjects' ratings could be attributed to the contrast effect. However, contrast effects in all other conditions (e.g., high-high-low) were only in the 1 - 2% of variance range.

Because Wexley et al. (1972) used artificially distinct categories of applicant quality ("completely qualified" and "completely unqualified"), Kopelman (1975) endeavored to narrow the range between high- and low-quality applicants. After testing for contrast effects with his more realistic applicants, Kopelman found that the effects actually accounted for only 11% of the variance in rating average applicants.

Schuh (1978) found that depending on which of the other applicants (if any) were seen before a target applicant, ratings could vary considerably, especially for average-quality applicants (as much as 58% in this study). Overall, the contrast effect accounted for almost as much variance as the applicant quality level (6.5% and 8.7%, respectively). In addition, contrast effects were shown to operate more strongly early in a series of applicants, whereas later in the series the contrast effect diminished. Therefore, Schuh recommends that applicants be seen in groups of more than four in order to give the contrast effect a chance to deteriorate. However, his recommendation also implies waiting until all applicants have been interviewed before assigning ratings, which is not likely to yield accurate assessments either due to difficulty in recall.

In related research, Heneman, Schwab, Huett, and Ford (1975) examined the effects of applicant order on interview validity. Students and job content experts viewed six videotaped interviews presented in one of two randomly selected orders and rated applicants according to their expected behavior on the job, should they be hired.

Supervisor ratings on the same job-performance rating instrument were used to validate the interview. Validity varied significantly depending on the order in which applicants were presented in the student interviewer group (average validity was .11 for the first applicant order and $-.13$ for the second applicant order). When validity was examined for the job content expert interviewers, interactions of applicant order by interviewer exposure to biodata about the applicant were present. Apparently, interviewer judgments of applicant suitability can vary in accuracy depending on when an applicant is seen in relation to others. Heneman et al. (1975) also mention in their discussion that interrater reliability did not seem to be influenced by the order in which applicants were presented, although no formal analyses were presented.

Willihnganz (1991) examined whether applicants might be rated differently by panel interviewers depending on their relative position to other applicants. Aggregate interview scores for 818 applicants were separated into six groups based on their interview slot (i.e., either first interview of the day, midmorning, just prior to lunch, just after lunch, midafternoon, or last interview of the day). No significant differences in interview scores according to relative position of applicants was found. Willihnganz speculated that perhaps the highly structured interviewing process used had prevented the emergence of differences according to slot.

Although there were no statistically significant differences in the scores according to relative position of the applicants, Willihnganz did discover a pattern in the ratings (no explanation for the emergence of this pattern was proposed). Interview scores rose throughout the morning, peaking just prior to lunch, then fell steadily again in the afternoon. Applicants were rated the lowest if they had the first interview of the day or the last interview of the day. This finding suggests that interview scores may vary

systematically according to where in the slate of applicants a particular applicant is seen.

The methodology of the Willihnganz study leaves the door open for more questions about applicant order to be answered. One is whether the slot the applicant occupies in the interview slate might have had an effect on interviewer evaluations if the evaluations had been operationalized differently. Willihnganz used a mechanical combination of ratings as the dependent variable of interview performance. However, many interviews utilize a subjective judgment of the applicant rather than a mechanical combination of ratings to make the final decision on an applicant. Could there be an effect of applicant order if interviewers use subjective judgments? In addition, the ratings used by Willihnganz were assigned after the interviewers had discussed the applicant (thus explaining the very high interrater reliability of .97). Might the applicants' relative position have affected ratings if interviewers had rated each applicant individually, without having an opportunity to be influenced by their fellow interviewers?

Purpose of This Study

The research presented here demonstrates that there is still much to be learned about factors that may moderate the effectiveness of multiple sequential panel interviews, particularly for the situational factor of applicant order. At the conclusion of their literature review, Arvey and Campion (1982) reported that interview research has neglected the investigation of "situational factors as they impact interviewers" (p. 311). Several years later, Eder (1989) concurred, lamenting how little is known about situational factors in the interview. This state of affairs is especially regrettable to the human resources practitioner, as situational factors are generally easier to manipulate and control, whereas applicant and interviewer factors (e.g., preparation, biases, decision-making processes, etc.) are either "givens" or are more difficult to change.

The Willihnganz (1991) study is the only one which has looked at applicant order in an actual employment setting. And, although Willihnganz did not find significant differences in scores according to the position of the applicant in the panel interview slate, there are other possibilities for exploring this phenomenon, as stated earlier. Although the laboratory research on context effects is interesting, it is very microanalytic in nature, and the conclusions lend themselves to few practical applications. For example, human resource practitioners cannot construct an interview slate so that preceding applicants are of similar quality in order to avoid a contrast effect. Therefore, the focus of this study is on actual interviewer evaluations.

Dependent variables in the research on applicant order have included ratings of job qualifications, ratings of interview performance, ratings on job-related dimensions, pass-fail decisions, and interview validity. Also, several moderators of interrater reliability in the panel interview have been investigated (structure, number of judgments made, etc.). However, with the exception of Heneman et al.'s (1975) informal analysis of data from a laboratory experiment, researchers have yet to examine whether applicant order might have an effect on interrater reliability. More needs to be learned about this topic given that sufficient reliability is a prerequisite to validity of the interview.

Hypothesis 1: Interviewers' interrater reliability varies depending on the relative position an applicant occupies in a series of panel interviews. Given the lack of research in this area, this hypothesis is nondirectional.

Differences in reliability and validity have been observed for different methods of arriving at a final interview decision (Conway et al., 1995; Wiesner & Cronshaw, 1988; Pulakos et al., 1996). If different methods of producing final interview outcomes can affect reliability and validity, they may also vary systematically depending on applicant

order. Willihnganz (1991) examined whether mechanically combined scores might vary according to the relative position of an applicant. Perhaps if interviewers were asked to give a more subjective evaluation of the applicant, differences in interview outcomes would result. Therefore, this study endeavors to examine each of two methods of producing final interview outcomes to determine whether one is more stable than the other in multiple sequential panel interviews.

Hypothesis 2: The value of interviewers' average ratings of applicants (mechanically combined) varies depending on the interview slot the applicant occupies in a series of panel interviews. Given the pattern of data Willihnganz (1991) presented, the ratings are expected to rise throughout the morning, come to a peak near the middle of the day, and then sink even lower than first interview levels by the end of the day.

Hypothesis 3: The likelihood of applicants' obtaining a favorable interview outcome (based on interviewers' subjective decisions about whether they would recommend an applicant for the position) varies depending on the interview slot the applicant occupies. Although Willihnganz did not examine subjective ratings, the likelihood of a positive recommendation is expected to follow the pattern for average ratings (i.e., rise throughout the morning, come to a peak near the middle of the day, and then sink even lower than first interview levels by the end of the day).

Knowing whether interrater reliability, interview ratings, and interview outcomes vary according to the relative position of applicants would be an important addition to the interviewing literature. As mentioned earlier, it is especially important to discover more about interview situational factors such as applicant order because of the potential ease of manipulating these factors. If effects due to applicant order exist, there are implications for the fairness of the panel interview process when multiple sequential interviews are used.

Knowing about these effects could prompt organizations to make adjustments in their scheduling of panel interviews (such as scheduling fewer interviews per day) or to engage in interviewer training to prevent errors in judgment resulting from order effects. Applicants could also use information about order to choose an optimal time for their interview.

Method

Overview

This study was based on records of panel interviews conducted in the Northern California division of a large health maintenance organization from August 1994 through October 1994. The interviews were part of the selection process for 37 newly created leadership positions (many with multiple openings). Only applicants whose applications demonstrated that they met the basic position qualifications were invited to an interview. The purpose of the interview was to provide additional information regarding applicants' suitability for the positions. Recommendations from the interviews were passed on to the senior leadership team of the organization, who considered the recommendation along with other available information (such as reference checks and performance appraisals) to decide who would go on to second interviews.

Archival data were available for almost all of the interviews ($N = 356$). Although the intent of the organization where the interviews were conducted was to store all interview documentation, apparently a small percentage of the data was discarded accidentally. Neither the interviewers nor the applicants were aware that data from the interviews would later be used for this study.

In order to be able to detect patterns in the data, several criteria were used to determine which interview records would be used in the analysis:

1. The data originally covered panel interviews with varied schedules; however, in order to avoid a confound by time of day, only data from panel interviews following the most prevalent schedule were selected for the analysis (five sequential 60-minute interviews with a 30-minute debriefing session between interviews).

2. Although all interviews were conducted by more than one interviewer, data were not available for every interviewer. Without having at least two rating forms, interrater reliability could not be calculated. Therefore, interviews for which there was only one rating form in the records were excluded.

3. Most of the interview days were “full” (i.e., each of the five interview time slots was filled with an applicant). However, in some cases not all slots were filled; also, data were sometimes missing for a particular applicant. In order to have the best chance of detecting effects if they were present, only records for full interview days were included in the analysis.

4. Most interviewers stayed for the entire set of five interviews. However, sometimes a “guest” interviewer would come in for one interview and not stay for the remainder, or perhaps an interviewer would leave early and so only be able to see four of the five applicants. Because the dynamics of the core group of interviewers were of interest, all interviewers who did not rate each of the five applicants were eliminated from the analysis.

After these four criteria were applied, data from 155 interviews conducted as part of the hiring process for 17 leadership positions remained for analysis. There were 31 interviews yielding 70 rating forms in each of the five interview slots, for a total of 350 rating forms.

Applicants

Data concerning the evaluation of 136 applicants were included in the analysis. There were more interviews than applicants because some applicants were interviewed separately for more than one position (i.e., 10% interviewed for two positions and 2% interviewed for three or four positions). In two cases, an applicant interviewed twice in the same time slot, but for different positions and with a different panel of interviewers.

Very little applicant demographic data were collected. Of the 131 applicants for whom information on sex was available, a slight majority (52%) were female. Although no official data were available regarding applicant age, personnel involved in administering the interviews estimated that applicants were mostly in their 30s and 40s. Informally, it was also observed that almost all of the applicants were internal to the organization.

Interviewers

Data from 54 interviewers were included in the analysis. Interviewers mainly consisted of upper level leadership and professional level job content experts. Efforts were made to include interviewers from minority groups and to include representatives from each of the three major entities of the organization. A majority of the interviewers (59%) were female. Again, no official data on age were available; however, those involved in administering the interviews estimated that interviewers were mostly in their 40s.

Interviewers were organized into 31 unique interview panels. Of the 31 panels, the majority (77%) consisted of two interviewers; 10% consisted of three interviewers, and 3% (a single panel) consisted of four interviewers. Interviewers could serve on more than one panel as long as they did not take place on the same day. The majority of the interviewers (78%) served on only one interview panel, 19% served on two interview panels, and 4%

served on four interview panels. There were 70 interviewers for each of the five interview time slots.

Interview Design

Dimensions. Despite the uniqueness of each of the positions' duties, all of the positions were alike in that they were considered to be "leadership positions." Therefore, applicants were assessed on the same 10 "key leadership competencies," or dimensions. These dimensions had been developed prior to the interviews by senior leadership in conjunction with internal human resources consultants, and were intended to be used eventually for selection, performance appraisal, and compensation. The 10 dimensions were (a) Customer Service, (b) Performance and Results Focus, (c) Strategic Thinking, (d) Staff Development and Empowerment, (e) Ability to Facilitate/Implement Change, (f) Collaboration and Team Stewardship, (g) Communication Skills, (h) Problem-Solving and Decision-Making Style, (i) Self-Learning and Development, and (j) Business and Technical Knowledge. Full definitions of each of the 10 dimensions are available in Appendix A (the interview rating form).

Questions. Using the 10 dimensions, written job descriptions, and (in some cases) consultations with hiring authorities, human resources personnel wrote interview questions for each of the positions. Because position duties were unique, there was little overlap among the questions for the 17 positions. Questions were either behavioral (in which applicants were asked to share past experiences that demonstrated their ability to perform the job) or situational (in which applicants were asked to imagine what they would do in hypothetical situations if they had been hired for the job). Questions typically contained multiple parts, with the first being the main question and the remainder being standard probes to use if the applicant did not supply enough information. In addition, each

question was followed by a notation specifying which of the 10 leadership dimensions was being tapped. For most positions, 8 - 12 questions were generated.

Rating forms. A standardized rating form (presented in Appendix A) was developed for the interviewers' use. The rating form asked interviewers to rate the degree to which each applicant exhibited behaviors or had knowledge/experience relevant to each of the 10 leadership dimensions using a 3-point rating scale, as follows:

1 = *Low: None or little demonstration of this competency*

2 = *Medium: Moderate demonstration of this competency*

3 = *High: A great deal of demonstration of this competency*

If the interviewer was unable to provide a rating on a given dimension, he or she was instructed to choose the response of *Don't Know* for that dimension. Interviewers were not asked to rate the applicant's response to each question; rather, they were asked to provide a global rating on each dimension. In addition, space was provided for interviewers to write comments regarding the applicant. The last part of the rating form asked whether the interviewer would recommend the applicant for the position, to which the interviewer could respond *No*, *Maybe*, or *Yes*.

Interview Administration

Overview. Applicants were scheduled for an interview time in rough alphabetical order by the Personnel Department. Each of the panels interviewed five applicants per day for 60 minutes each, with a 30-minute evaluation period following each interview. The panels' schedule was as follows:

8:30 a.m. - 9:00 a.m.	Interviewer orientation
9:00 a.m. - 10:00 a.m.	Interview of first applicant (Slot 1)
10:00 a.m. - 10:30 a.m.	Evaluation of first applicant

10:30 a.m. - 11:30 a.m.	Interview of second applicant (Slot 2)
11:30 a.m. - 12:00 p.m.	Evaluation of second applicant
12:00 p.m. - 1:00 p.m.	Lunch
1:00 p.m. - 2:00 p.m.	Interview of third applicant (Slot 3)
2:00 p.m. - 2:30 p.m.	Evaluation of third applicant
2:30 p.m. - 3:30 p.m.	Interview of fourth applicant (Slot 4)
3:30 p.m. - 4:00 p.m.	Evaluation of fourth applicant
4:00 p.m. - 5:00 p.m.	Interview of fifth applicant (Slot 5)
5:00 p.m. - 5:30 p.m.	Evaluation of fifth applicant
5:30 p.m. - 6:00 p.m.	Final decisions

Most of the interview panels (84%) conducted interviews for only one position the entire day. Another 13% of panels conducted interviews for two positions, and 3% conducted interviews for three positions. In all cases of panels conducting interviews for more than one position, the majority of the interviews (three or four) were for one position, and only one or two were for other positions.

Interviewer preparation. Prior to the day of the interviews, interviewers were sent applications and performance appraisals of the applicants they were to see in order to familiarize themselves with the applicants. They were also sent a schedule, a memo informing them of the purpose and structure of the interview, a job description, a copy of the interview questions, and a sample rating form.

Upon arrival at the interview site, interviewers were greeted by a facilitator provided by the Personnel Department who oriented them to the interview process and distributed materials. The orientation consisted of introductions, a review of the interview purpose and structure, a review of the applicant slate, and deciding who would ask which questions of

the applicants. There was no interviewer training beyond the 30-minute orientation.

Interview. Interviewers asked applicants the same mandatory core questions for each position in the same order. When necessary, the prespecified probes provided for each question were used to elicit more information from the applicant. If time allowed, interviewers could choose to ask two to three prespecified optional questions, which were also the same for all applicants for a given position. Interviewers decided in advance who would ask applicants each question; generally interviewers took turns and asked two to five questions over the course of the interview. The same question-asking order was maintained throughout the day. Facilitators were not present during the interviews, only during the evaluation session following each interview.

Evaluation. Immediately following each applicant's interview, interviewers had 30 minutes in which to evaluate the applicant. First, interviewers completed rating forms separately. Only after completing their individual forms were interviewers allowed to discuss their ratings, comments, and recommendations. After the discussion period, the facilitator asked each interviewer to give his or her final ratings of the applicant. The facilitator recorded these final ratings on a separate form. Although interviewers were not asked to come to consensus about their final ratings on the 10 dimensions, they were strongly encouraged to reach consensus on their final recommendation of the applicant for the position. After final ratings were recorded, the facilitator collected the interviewers' individual rating forms.

Final decisions. At the end of the day, the facilitator confirmed the panel's final recommendation for each applicant (i.e., *No*, *Maybe*, or *Yes*). The interviewers were not asked to compare the applicants to each other or to rank them.

Results

Dependent Variables

There were two dependent variables used to analyze results, recommendation and average score. Recommendation was a categorical variable with three values: *No*, *Maybe*, and *Yes*. This variable represented individual interviewers' decisions about whether they would recommend the applicant for the position.

Average score, the second dependent variable, was obtained by taking the average of 8 of the 10 dimension ratings provided by each interviewer. Of the original 10 dimensions, 8 had less than 5% missing data, but 2 (Staff Development and Empowerment, and Self-Learning and Development) had 23% and 14% missing data, respectively. Therefore, the decision was made to exclude these two dimensions when calculating average score.

If an interviewer did not supply a rating on one of the eight dimensions, average score was computed using the dimension values that were rated. The majority of interviewers (83%) were not missing any ratings and 13% were missing only one rating. The smallest number of ratings on which the average score was based was four (one interviewer).

Interrater Reliability (Hypothesis 1)

For both of the reliability analyses, there were not enough three-person and four-person panels to use in calculating the statistics. Therefore, data from only two randomly selected interviewers from each of the three- and four-person panels were used in arriving at the reliability statistics.

Intraclass correlations were used to test the interrater reliability of the average score variable. Table 1 shows the intraclass correlations of average score for each slot. One-way

Table 1

Interrater Reliability of Average Score

Slot	Intraclass correlation	F_o	Between groups df	Within groups df
1	.73	6.37	30	31
2	.83	10.51	30	31
3	.75	6.97	29	30
4	.64	4.61	29	30
5	.87	14.72	30	31

analyses of variance (ANOVAs) testing whether the agreement among interviewers was significantly greater than 0 in each slot were all significant at the $p < .05$ level (see Shrout & Fleiss, 1979). The last slot appeared to have the highest interrater reliability, whereas the other four slots had lower interrater reliability. Given these results, a decision was made to test the Slot 1 correlation (first interview) against the Slot 5 correlation (last interview) using the z significance test for two independent correlations (see Lehman, 1995, p. 294). The difference between the correlation for Slot 1 and the correlation for Slot 5 was nonsignificant ($z = 1.51, p > .05$).

To test the reliability of the recommendation variable, kappa analyses were conducted. Kappa statistics of recommendation for each slot are presented in Table 2. As in the tests for reliability of average score, tests to determine whether agreement among interviewers was significantly greater than 0 in each slot were all significant at the $p < .05$ level (see Fleiss, 1971; Hubert, 1977). Mirroring the average score reliability results, the last slot once again had the highest interrater reliability (in fact, perfect agreement), whereas the other four slots had lower interrater reliability. Again, the Slot 1 correlation was tested against the Slot 5 correlation using the z significance test. The results of the analysis indicated that the Slot 5 correlation was significantly greater than the Slot 1 correlation ($z = 5.15, p < .05$).

Mechanically Combined Ratings (Hypothesis 2)

In order to determine whether interviewer evaluations as measured by average score varied according to slot, a one-way within-subjects ANOVA was conducted. This analysis was appropriate given that each interviewer was given five rating tasks in the course of his or her interviewing day (a repeated-measures design). Table 3 presents the means and standard deviations of the average score for each slot; Figure 1 graphs the

Table 2

Interrater Reliability of Recommendation

Slot	Kappa	<i>T</i>	<i>n</i> ^a
1	.75	4.78	23
2	.85	5.53	22
3	.74	4.42	19
4	.62	3.97	21
5	1.00	5.19	21

^apairs of interviewers

Table 3

Average Score for Each Interview Slot

Slot ^a	<i>M</i>	<i>SD</i>
1	2.37	.47
2	2.43	.46
3	2.38	.50
4	2.38	.54
5	2.33	.54
All	2.38	.50

^a*n* = 70 rating forms for each slot.

Figure 1. Average score for each interview slot.

means by slot. Contrary to predictions, there were no differences in average score depending on the interview slot, $F(4, 276) = 0.32, p > .05$.

Subjective Recommendations (Hypothesis 3)

In order to learn whether interviewers' likelihood of recommending an applicant positively or negatively varied significantly overall based on the relative position of the applicant, a chi-square analysis was performed on recommendation by slot. Differences in interview outcomes according to the slot were not significant overall, although they did approach significance at the $p < .05$ level, $\chi^2(8) = 13.91, p = .08$. However, as shown in Table 4 and Figure 2, there did appear to be a definite pattern of recommendation depending on the slot the applicant occupied. It appeared that interviewers become increasingly less likely to give a positive recommendation and more likely to give a negative recommendation as their day went on. Placements in the *Maybe* category also declined as the sequence of applicants progressed. Applicants in the first slot appeared to have the greatest chance of being recommended positively. Those seen in Slots 2, 3, and 4 seemed to have very similar chances of being recommended positively. Finally, applicants who were seen in the fifth slot seemed to be less likely to receive a positive recommendation.

Based on the trends in the data and the fact that the p value in the initial chi-square approached significance, a second chi-square analysis was conducted to test whether recommendation differed according to membership in three slot subgroups which follow the groupings suggested by Table 4 and Figure 2. The three slot groups were (a) first slot, containing Slot 1; (b) middle slots, containing Slots 2, 3, and 4; and (c) last slot, containing Slot 5. When the data were grouped in this manner (presented in Table 5), significant

Table 4

Number and Percentage of Recommendations for Each Interview Slot

Slot	No	Maybe	Yes	Total
1	10	14	37	61
	16%	23%	61%	100%
2	18	10	30	58
	31%	17%	52%	100%
3	15	13	30	58
	26%	22%	52%	100%
4	17	8	34	59
	29%	14%	58%	100%
5	26	6	27	59
	44%	10%	46%	100%
All	86	51	158	295
	100%	100%	100%	100%

Figure 2. Recommendations for each interview slot.

Table 5

Number and Percentage of Recommendations for Each Interview Slot Group

Slot Group	No	Maybe	Yes	Total
First slot (1)	10	14	37	61
	16%	23%	61%	100%
Middle slots (2, 3, 4)	50	31	94	175
	29%	18%	54%	100%
Last slot (5)	26	6	27	59
	44%	10%	46%	100%
All	86	51	158	295
	100%	100%	100%	100%

differences in recommendation according to slot group at the $p < .05$ level were discovered. $\chi^2(4) = 12.05, p = .02$. From examining the pattern of the recommendations, it appears that applicants assigned to the first slot had the best chance of being recommended positively (only 17% were rejected). Applicants assigned to middle slots were less likely to be recommended positively (29% were rejected). Finally, the applicants assigned to the last slot had the worst chance of being recommended positively (44% were rejected).

Discussion

The purpose of this study was to investigate the effect of an applicant's relative position in a slate of sequential panel interviews on interrater reliability, the value of the interview ratings the applicant received, and the likelihood that the applicant would obtain a favorable interview outcome. This section will discuss the results of the study and their implications, as well as delineating limitations of the study and possibilities for future research on applicant order.

The first hypothesis, that interviewers' interrater reliability varies depending on the interview slot the applicant occupies, was partially supported. Reliability was not identical across slots for both the average score and recommendation dependent variables. Differences were most apparent for the last interview slot, which tended to have higher reliability coefficients. However, when the difference in reliability of the fifth slot was tested against the first slot, only the recommendation variable showed significantly higher interrater reliability for the fifth slot.

The second hypothesis, that interviewers' mechanically combined ratings vary depending on interview slot, was not supported. There were no significant differences in average score by slot. However, the pattern of the data was similar to the one observed by

Willihnganz (1991), who also found no significant differences according to relative position of the applicants. The first slot had slightly lower average ratings than interviews in the middle of the day, and the last slot had the lowest ratings of all.

The final hypothesis, that interviewers' subjective decisions (recommendation) about applicants vary depending on the interview slot the applicant occupies, was partially supported. Overall there were no significant differences by slot, although the test did approach significance ($p = .08$). However, a post hoc analysis comparing subgroups of the original five slots (first slot, containing Slot 1; middle slots, containing Slots 2, 3, and 4; and last slot, containing Slot 5) revealed significant differences. Applicants who were assigned to the first slot seemed to have the greatest likelihood of being recommended positively, followed by applicants in the middle slots. Applicants who had the misfortune of occupying the last slot had the least chance of being recommended positively. Although it was predicted that recommendations would follow the pattern reported by Willihnganz (1991), they did not; a greater proportion of first slot applicants received positive recommendations than were expected.

Explanation of Results

The most striking finding in this study is that the relative position of applicants produced different effects on the two types of interviewer evaluations. That is, average scores across the slots were not significantly different, whereas interviewer recommendations were different.

One possible explanation for these findings is that the rating instrument provided a more structured evaluation exercise for interviewers than did the request for a subjective recommendation. In explaining his findings, a lack of differences in ratings according to slot, Willihnganz (1991) suggested that the highly structured interview process used may

have controlled outside factors which otherwise would have influenced ratings. Indeed, using a rating form is part of the structured interview process (Campion et al., 1988), which is known to have higher reliability and validity than unstructured interviews. When interviewers were faced with the subjective evaluation task, other factors which had been controlled successfully by the structured rating exercise may have had a chance to operate.

One of the outside factors that may have been able to exert an influence is fatigue. It is certainly reasonable to assume that fatigue may have been a side effect of interviewing a series of applicants under time pressures. The fatigue explanation was suggested by anecdotal information supplied by the interviewers themselves, whose most frequent complaint was that five interviews in one day were too many. They admitted they “could not do the last applicant justice.”

Fatigued interviewers may have had less tolerance for average-quality applicants at the end of the day than they did at the beginning of the day, when they were well-rested and perhaps in a better mood. In support of this explanation, Baron (1987) found that interviewers who were placed in a negative mood recalled more negative information about applicants and made less favorable employment decisions about them. Furthermore, research has shown that interviewers generally pay more attention to negative information about applicants than they do positive information (e.g., Bolster & Springbett, 1961; Hollman, 1972). Perhaps this tendency was exaggerated either by fatigue or by the lack of a structured evaluation exercise, resulting in interviewers’ placing more applicants in the *No* category.

Fatigue may also have interacted with the rating instrument’s effectiveness in providing structure (and thereby controlling outside influences such as personal bias). Perhaps earlier in the day interviewers were more likely to make their recommendations

conform to decisions they made on the rating form. Later in the day they may have tended to disregard their earlier decisions as recorded in the dimension ratings and rely more on “gut feelings” about the applicant when arriving at a final recommendation. In addition, fatigue may have decreased interviewers’ ability to execute successfully the cognitively complex task of integrating all the information about the applicant to arrive at an overall decision. Conway et al. (1995) allude to this in their finding that interviewers who combine ratings subjectively to arrive at a decision make poorer decisions than those who combine ratings mechanically.

Aside from the degree of structure built into the rating task and fatigue, another possible explanation for the results is that interviewers were reacting to perceived pressures to build an adequate applicant pool. Although interviewers did not have a hiring quota, they were aware of the numerous openings for each position, and the fact that there was a second interview ahead that would further narrow the pool. Perhaps interviewers were more lenient in their recommendations at the beginning of the day in an attempt to ensure an adequate applicant pool. By the end of the day, even though applicant quality was no different (as assessed by the average score), interviewers may have believed enough applicants had been forwarded to the second interview and decided there was no longer a need to forward average-quality applicants.

Yet another explanation for the order effect is that the interviewers may have become more decisive by the end of the day due to their experience. Earlier they may not have felt so confident about how their ratings should relate to their final recommendation about an applicant. By the end of the day, they may have had enough exposure to applicants of varying qualities, exposure to other interviewers’ opinions, and opportunities to engage in discussions about applicants that they had a better grasp about what it would

take for an applicant to be successful in the position. Thus their *perceptions* of the meanings of the ratings may have changed and been reflected in their recommendations, although the *value* of the ratings had not changed.

Because this study was not experimental, an important question about causality should be raised. Was it interviewers or applicants who were different at the end of the day? Perhaps applicants participating in the interviews were freshest in the morning, before they had worked a full day (almost all of the applicants were internal to the organization). This might provide an explanation for the small, nonsignificant dip in ratings seen in both this study and in the Willihnganz (1991) study. However, the magnitude of the differences in recommendation were so much greater than warranted by the nonsignificant patterns in average score that explanations should center, rightly, on interviewer factors.

The results of the reliability analysis are more difficult to speculate about. Interviewers tended to agree the most on the last interview and less on the previous four. On the subjective recommendation, they agreed significantly more for the last applicant than for the first. Perhaps interviewers were striving to agree with each other throughout the interviewing process and eventually were able to make their ratings converge as they got to know their fellow interviewers better. This kind of effect might be especially strong in situations where higher ranking interviewers were matched with lower ranking interviewers. The lower ranking interviewers may have made an effort to anticipate what the other interviewer would recommend.

Interviewers' use of the rating instrument could also have resulted in changes in reliability. For example, interrater reliability could be higher for the last slot because interviewers were making more rating errors due to fatigue. The increased presence of

certain rating errors, such as halo or central tendency, could provide an explanation for increased agreement. Another explanation is that by the end of the day, interviewers may have relaxed their adherence to the interview guidelines, perhaps sharing their impressions and overall recommendation with each other prior to completing their rating instrument. If this was the case for even a few interviews, it would certainly enhance agreement.

Practical Implications

Because fairness to applicants appears to be reduced by allowing interviewers to make subjective overall recommendations, the most obvious implication of this study for the interview process is to allow only mechanically combined ratings for evaluating applicants. The average score provided by interviewers in this study did not seem to be swayed by the slot an applicant occupied.

Another possibility is to reduce the toll the interview process takes on interviewers. One way of doing this is to limit the number of sequential interviews that interviewers may conduct. Perhaps restricting the number of interviews to four (because effects of order were the worst for the fifth applicant) would have resulted in fairer outcomes for all applicants, regardless of what time they were seen. Building in rest time and breaks, regardless of the number of interviews conducted, might also improve the ability of interviewers to be consistent in their evaluations of applicants.

Yet another possibility is to provide opportunities for interviewers to practice the task they have been asked to complete. Although they were oriented to what was expected of them, the interviewers in this study received no training on rater biases and errors and had no practice in using the questions or rating form until they began interviewing. Perhaps greater familiarity with the evaluation task would have improved interviewers' ability to provide accurate assessments. Prior research has shown that training can

improve the quality of interviewer decisions (Latham, Wexley, & Pursell, 1975).

Paying attention to panel composition might also improve fairness to applicants across interview slots. The interviewers in this study were not very familiar with the position, as it was new, and had not (for the most part) participated in constructing interview questions. Perhaps having more knowledgeable, involved interviewers would lead to less of a need to learn “on the job” (i.e., while interviewing), resulting in more consistent evaluations across slots.

Finally, applicants who are participating in sequential panel interviews should be aware of the effect their position in the panel could have on their evaluation. Given the choice of an interviewing time, applicants may want to select an earlier slot and avoid the last slot of the day.

Strengths and Limitations

A strength of this study is that actual employment interviews were used rather than simulated interviews in the laboratory, thereby increasing its generalizability. In addition, the sample size was relatively large, allowing a winnowing of the cases to eliminate possible confounds (such as absolute time of day, “guest” interviewers who might change the interview dynamics, fewer than five interviews conducted, etc.). This study was also able to examine interviewer opinions prior to the group discussion period, yielding more information about decision-making processes used by individual interviewers in the panel interview than prior research (i.e., Willihnganz, 1991).

However, there are some limitations. One is interviewers’ lack of familiarity with the position; what they knew about the position was based only on a written job description. This may have resulted in difficulty deciding how the ratings might translate into an overall recommendation, allowing bias to creep in. Another weakness is the fact

that the interviews were for many different positions, using different interview questions. Perhaps differences in average score may have been apparent if the interviews had been conducted for only one position.

Another limitation is the restrictive 3-point scale interviewers used for assigning dimension ratings. If there had been more differentiation available in the ratings (by using a 5- or 7-point scale, for example), differences that may exist for average score according to slot would have had a better chance of emerging. Additionally, using the average score as the dependent variable could have masked differences according to order that may have been apparent upon closer examination of each individual dimension.

Regarding interrater reliability, the fact that only two interviewers were used for the reliability calculations may have decreased the study's ability to detect systematic differences in reliability by slot. Finally, the study is limited in that no information is available about interviewers' state of mind (e.g., fatigue, understanding of position requirements, etc.) as they participated in the rating task. Therefore, explanations for the effects shown must rely on speculation.

Future Research

This study provides many avenues for future research. One of the primary areas of interest is to explore the limits of the order effect reported in this study. Do interviewers give harsher judgments after conducting a certain number of interviews per day (in this study, four)? Or does the last applicant receive harsher judgments no matter how many others were seen before? Courses of action for human resources practitioners vary depending on the answer to these questions. For example, if the last applicant always receives the harshest treatment, training interviewers or using only mechanically combined ratings might be the most appropriate intervention. Alternatively, if applicants are treated

fairly when seen in line-ups of three or fewer, a logical solution would be to limit to three the number of applicants interviewers may see in one day.

Drawing from a limitation of this study mentioned earlier, researchers might also want to replicate this study using a different scale (e.g., a 5- or 7-point scale) to allow more variability in interviewer ratings. If average scores are at all affected by interview slot, this change in instrumentation would be more likely to uncover them than the 3-point scale used in this study. Therefore, if findings were similar to those reached in this study, it would add more strength to the conclusion that mechanically combined ratings are fairer barometers of applicant quality than subjective recommendations. However, if results using a revised scale show effects due to applicants' position in the interview slate, it would suggest that the order effect is more pervasive than this study indicates.

Future researchers should consider investigating causes of the order effect, such as fatigue. Having interviewers provide a measure of their fatigue immediately after each interview would allow any trends to emerge that may be related to the differences in recommendation. For example, fatigue may increase as likelihood of being recommended positively decreases, as suggested in this study. As several factors may be operating at once to produce the order effect (e.g., fatigue, understanding of position requirements, perceived pressures to hire, etc.), it may also be useful to look at fatigue in conjunction with other possible causes in a single study to determine how much each factor contributes to the effect.

Interviewer preparation is another area that might be fruitfully examined in future research. Because interviewer training has been shown to mitigate some types of errors in judgment (Latham et al., 1975), training interviewers might eliminate the order effect observed with subjective ratings. In addition, interviewers might be better prepared for the

recommendation task by virtue of previous experience with the interview process and rating instrument. Examining whether more prepared interviewers also exhibit differences in recommending applicants according to interview slot would provide more insight into factors contributing to this effect.

As discussed earlier, the results suggest that interviewers may have been influenced by perceived hiring quotas as they evaluated applicants. It could be that as interviewers sense they have contributed enough applicants to the pool, they are not as likely to forward applicants who are of average quality. To test this idea, future researchers could manipulate interviewer expectations about the state of the applicant pool to investigate how these expectations might mitigate the order effect seen in this study. For example, the manipulation could be accomplished by telling one interviewer group that 50 applicants are interviewing for five openings and telling another group that 10 applicants are interviewing for five openings. If interviewers are influenced by the state of the applicant pool, the order effect should be more apparent for the group who believes the pool is smaller.

Another area for future research to examine is how interview slot might affect a panel's group decision as opposed to interviewers' individual opinions. Recall that the ratings and recommendations used in this study were interviewers' opinions *prior* to group discussion. If group decisions are the final decisions, individual interviewers' errors due to applicant order may not be important. Some have speculated that the discussion period is the most beneficial part of panel interviews (Dipboye, 1992; Gatewood, 1990; Weston & Warmke, 1988). Perhaps order effects can be moderated or eliminated by the process of interviewers' sharing their impressions and discussing the applicant.

Finally, it is important for future researchers to validate interviewer decisions according to interview slot by collecting interview ratings for those who were hired and

later comparing them to performance ratings. This would help answer the question: Do interviewers make higher quality decisions earlier in the day or later in the day? As discussed above, the interviewers' tendency to turn down applicants seen in the last slot may be a reflection of their increased understanding of what the needs are for the position. If this is the case, interviewers actually may have done a better job in predicting who would perform better on the job for the last applicant. Furthermore, reliability was higher for the last interview, increasing the chances that validity is also higher. The only way to tell whether the last applicant was judged more appropriately is to do a validity study.

Conclusion

This study has demonstrated that applicant order can have a dramatic impact on interview outcomes when interviewers are asked to make subjective recommendations about applicants. This is especially true for the last applicant seen in a day, who is less likely to receive a positive recommendation. Fortunately, mechanically combined ratings seem to be immune to the order effect. However, before a course of action to control for the order effect can be recommended, there is still much to be learned about why and under what circumstances the order effect occurs. Undertaking further research as suggested in this study is vitally important not only to maximize fairness to applicants undergoing evaluation in the panel interview, but also to improve the usefulness of the panel interview to organizations.

References

- Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology*, 35, 281-322.
- Bamford, J. (1986, March 24). Surviving the group interview. *Forbes*, 137(6), 190-191.
- Baron, R. A. (1987). Interviewer's moods and reactions to job applicants: The influence of affective states on applied social judgments. *Journal of Applied Social Psychology*, 17, 911-926.
- Bell, A. H. (1989). *The complete manager's guide to interviewing: How to hire the best*. Homewood, IL: Dow Jones-Irwin.
- Bolster, B. I., & Springbett, B. M. (1961). The reaction of interviewers to favorable and unfavorable information. *Journal of Applied Psychology*, 45, 97-103.
- Bureau of National Affairs. (1988). *Personnel policies forum survey no. 146: Recruiting and selection procedures*. Washington, D.C.: Author.
- 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.
- Church, A. H. (1996). From both sides now: The employee interview--The great pretender. *The Industrial-Organizational Psychologist*, 34 (1), 108-117.
- Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology*, 80, 565-579.
- Cox, J. A., Schleuter, D. W., Moore, K. K., & Sullivan, D. (1989). A look behind corporate doors. *Personnel Administrator*, 34(3), 56-59.
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34, 571-582.
- Dipboye, R. L. (1992). *Selection interviews: Process perspectives*. Cincinnati, OH: South-Western.
- Eder, R. W. (1989). Contextual effects on interview decisions. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 113-126). Newbury Park, CA: Sage.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378-382.

- Gatewood, R. D. (1990). *Human resource selection* (2nd ed.). Orlando, FL: The Dryden Press.
- Heneman, H. G., III, Schwab, D. P., Huett, D. L., & Ford, J. J. (1975). Interviewer validity as a function of interview structure, biographical data, and interviewee order. *Journal of Applied Psychology*, 60, 748-753.
- Hollman, T. D. (1972). Employment interviewers' errors in processing positive and negative information. *Journal of Applied Psychology*, 56, 130-134.
- Hubert, L. (1977). Kappa revisited. *Psychological Bulletin*, 84, 289-297.
- Huffcutt, A. I., & Arthur, W., Jr. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology*, 79, 184-190.
- Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, 96, 72-98.
- Kopelman, M. D. (1975). The contrast effect in the selection interview. *British Journal of Educational Psychology*, 45, 333-336.
- Latham, G. P., Wexley, K. N., & Purcell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology*, 60, 550-555.
- Lehman, R. S. (1995). *Statistics in the behavioral sciences*. Pacific Grove, CA: Brooks/Cole.
- Mayfield, E. C. (1964). The selection interview: A reevaluation of published research. *Personnel Psychology*, 17, 239-260.
- 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.
- Pulakos, E. D., Schmitt, N., Whitney, D., & Smith, M. (1996). Individual differences in interviewer ratings: The impact of standardization, consensus discussion, and sampling error on the validity of a structured interview. *Personnel Psychology*, 49, 85-102.
- Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology*, 35, 1-62.
- Schmitt, N. (1976). Social and situational determinants of interview decisions: Implications for the employment interview. *Personnel Psychology*, 29, 79-101.
- Schuh, A. J. (1978). Contrast effect in the interview. *Bulletin of the Psychonomic Society*, 11, 195-196.

- Schwab, D. P., & Heneman, H. G., III. (1969). Relationship between interview structure and interinterviewer reliability in an employment situation. *Journal of Applied Psychology*, 53, 214-217.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420-428.
- Ulrich, L., & Trumbo, D. (1965). The selection interview since 1949. *Psychological Bulletin*, 63, 100-116.
- Wagner, R. (1949). The employment interview: A critical summary. *Personnel Psychology*, 2, 17-46.
- Warmke, D. L., & Weston, D. J. (1992). Success dispels myths about panel interviewing. *Personnel Journal*, 71(4), 120-126.
- Weston, D. J., & Warmke, D. L. (1988). Dispelling the myths about panel interviews. *Personnel Administrator*, 33(5), 109-111.
- Wexley, K. N., Yukl, G. A., Kovacs, S. Z., & Sanders, R. E. (1972). Importance of contrast effect in employment interviews. *Journal of Applied Psychology*, 56, 45-48.
- 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.
- Willihnganz, M. A. (1991). *Effects of time of day on interview performance*. Unpublished master's thesis, California State University, Sacramento.
- Wright, P. M., Lichtenfels, P. A., & Pursell, E. D. (1989). The structured interview: Additional studies and a meta-analysis. *Journal of Occupational Psychology*, 62, 191-199.

Initial Panel Interview Candidate Rating Form

Date: _____

Candidate: _____

Panelist: _____

- 1) For each of the following competencies, please rate the degree to which the candidate exhibited behaviors or has knowledge/experience relevant to that competency. If you are unable to provide a rating on a given competency, please circle "DK" for "Don't Know".
- 2) Under each of the competencies, please write in specific behavioral examples exemplifying the candidate's strengths and weaknesses on the competency to justify the rating you gave.

Low
None or little demonstration
of this competency

Medium
Moderate demonstration
of this competency

High
A great deal of demonstration
of this competency

Low Medium High

DK 1 2 3

Customer Service

Demonstrates a strong customer focus in all business decisions/actions. Demonstrates success in achieving high levels of customer satisfaction. Achieves continual improvements in service delivery in partnership with identified customers.

Strengths	Weaknesses

DK 1 2 3

Performance & Results Focus

Maintains a strong focus on ensuring all work/ organization activities are aligned to and facilitative of accomplishing key business goals. Consistently delivers timely and high quality results. Holds self and others to high standards of performance, and accountable for results achieved or not achieved.

Strengths	Weaknesses

Low Medium High
DK 1 2 3

Strategic Thinking

Consistently applies a customer-focused, integrative, boundary-free view to organization/management activities/issues. Attends to the longer-term implications of decisions/actions. Demonstrates a thorough understanding of the interdependencies of our organization's key business processes and management systems.

Strengths	Weaknesses
-----------	------------

DK 1 2 3

Staff Development & Empowerment

Facilitates the ongoing learning, well-being, and development of staff. Enables staff to achieve results by providing information, encouragement, support, and the discretionary authority to achieve agreed-upon goals, and meet customer needs. Demonstrates ability to assess the capabilities of others and provide work assignments that stretch those capabilities.

Strengths	Weaknesses
-----------	------------

DK 1 2 3

Ability to Facilitate/Implement Change

Actively challenges the status quo and traditional ways of working and assumptions. Demonstrates comfort and tolerance for uncertainty and ill-defined tasks, as well as a willingness to adjust planned approaches as situational demands shift. Has demonstrated significant performance improvements through the implementation of planned change or redesign efforts.

Strengths	Weaknesses
-----------	------------

	Low	Medium	High
DK	1	2	3

Collaboration & Team Stewardship

Able to effectively create and facilitate collaboration and cooperation among diverse groups, people, and professional disciplines to achieve maximum results. Able to gain cooperation and support through effective use of influence/persuasion. Demonstrates ability to recruit and select team members who possess complementary perspectives and skills relative to the task at hand. Has demonstrated a willingness to place own interests subordinate to the collective interests of a team. Track record of actively contributing as a member of a working team to achieving team results. Has achieved results by forming cooperative ventures with individuals/groups from diverse backgrounds and perspectives.

Strengths	Weaknesses

	Low	Medium	High
DK	1	2	3

Communication Skills

Presents ideas and information in a clear, concise manner both written and verbal. Able to quickly adapt communication style to different audiences. Demonstrates attention to and actively strives for understanding of others' ideas/opinions. Encourages the open expression of diverse ideas/opinions. Keeps others well-informed, and is open and honest with information.

Strengths	Weaknesses

	Low	Medium	High
DK	1	2	3

Problem-Solving & Decision-Making Style

Able to facilitate and/or make effective, timely decisions and resolve identified problems. Actively seeks out and includes the input of appropriate sources of expertise in decision-making/problem-solving. Identifies the root causes of problems, rather than simply addressing "symptoms," and examines a range of options before deciding on a course of action. Strives to "look outside the boxes" in analyzing problems/issues and developing solutions.

Strengths	Weaknesses

Low Medium High
DK 1 2 3

Self-Learning & Development

Seeks out opportunities for own learning and ongoing professional development. Actively solicits and is receptive to feedback from others, and shows an openness to views and ideas different from own. Shows an awareness of own strengths and development needs. Demonstrates ability to learn quickly from experience and feedback.

Strengths	Weaknesses

DK 1 2 3

Business and Technical Knowledge

Demonstrates a general management perspective and broad understanding of economic, social, and political trends and developments that may impact the business. Demonstrates a thorough and broad business and technical knowledge of our products, services, and business processes. Track record of effectively using sound financial indices to manage and guide business processes. *Demonstrates specific business and technical knowledge as specified in the PCQ.*

Strengths	Weaknesses

Please record any other perceptions of the candidate's strengths and weaknesses here. Please provide examples of actual behavior.

STRENGTHS	WEAKNESSES

Based on the information available to you (e.g., from the interview, portfolio, and screening information) would you recommend this candidate for the (name of) position?

☐ Yes; I would recommend this candidate for the (name of) position

☐ No; I would not recommend this candidate for the (name of) position

☐ Maybe; I would recommend this candidate for the (name of) position with the following reservations:

INTERVIEWER'S NOTES

San José State
UNIVERSITY

**Office of the Academic
Vice President**
**Associate Vice President
Graduate Studies and Research**

One Washington Square
San Jose, CA 95192-0025
Voice: 408-924-2480
Fax: 408-924-2477
E-mail: gstudios@whoo.sjsu.edu
<http://www.sjsu.edu>

TO: Susan Howard
3411 Brookmill Ct.
Fremont, CA 94536

FROM: Serena W. Stanford *Serena W. Stanford*
AAVP, Graduate Studies & Research

DATE: March 3, 1997

The Human Subjects-Institutional Review Board has approved your request to use human subjects in the study entitled:

"The Effects of Interviewee Order
in Panel Interviews"

This approval is contingent upon the subjects participating in your research project being appropriately protected from risk. This includes the protection of the anonymity of the subjects' identity when they participate in your research project, and with regard to any and all data that may be collected from the subjects. The Board's approval includes continued monitoring of your research by the Board to assure that the subjects are being adequately and properly protected from such risks. If at any time a subject becomes injured or complains of injury, you must notify Serena Stanford, Ph.D., immediately. Injury includes but is not limited to bodily harm, psychological trauma and release of potentially damaging personal information.

Please also be advised that all subjects need to be fully informed and aware that their participation in your research project is voluntary, and that he or she may withdraw from the project at any time. Further, a subject's participation, refusal to participate, or withdrawal will not affect any services the subject is receiving or will receive at the institution in which the research is being conducted.

If you have any questions, please contact me at (408) 924-2480.

The California State University:
Chancellor's Office
Bakersfield, Chico, Dominguez Hills,
Fresno, Fullerton, Hayward, Humboldt,
Long Beach, Los Angeles, Maritime Academy,
Monterey Bay, Northridge, Pomona,
Sacramento, San Bernardino, San Diego,
San Francisco, San Jose, San Luis Obispo,
San Marcos, Sonoma, Stanislaus