DISSERTATION

INTERVIEWER ACCURACY ACROSS LEVELS OF STRUCTURE IN THE EMPLOYMENT INTERVIEW

Submitted by

Elisa George

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2006

COLORADO STATE UNIVERSITY

March 13, 2006

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY ELISA GEORGE ENTITLED INTERVIEWER ACCURACY ACROSS LEVELS OF STRUCTURE IN THE EMPLOYMENT INTERVIEW BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Michael Gross

Jacob E. Hautaluoma

Randall Craig Swaim

Advisor – Eric Heggestad

Department Head – Ernest L. Chavez

ABSTRACT OF DISSERTATION

INTERVIEWER ACCURACY ACROSS LEVELS OF STRUCTURE IN THE EMPLOYMENT INTERVIEW

It is now common knowledge that more structured interviews result in greater criterion validity with respect to predicting performance. Theorists and researchers have also noted that interviewers differ in their ability to judge others. This study examined the relationships between interview structure, observer intelligence, and observer concept of the ideal candidate and observer accuracy. This study proposed that: 1) Greater interview structure would lead to greater observer accuracy; 2) General intelligence would moderate the relationship between interview structure and accuracy; and 3) Observers' concept of the ideal candidate would moderate the relationship between interview structure and accuracy. Three hundred and one undergraduate psychology students viewed mock employment interviews conducted with attitudinal questions, situational questions, and past-behavioral questions and rated candidate responses, using a behaviorally-anchored rating scale. Only the first hypothesis received limited support. Accuracy was significantly different among three interview conditions: an attitudinal interview (AI), situational interview (SI), and patterned behavioral description interview (PBDI). Accuracy was greatest in the SI, followed by the AI, and least in the PBDI. A small significant positive effect was found between intelligence and accuracy, and a

small negative effect was found between dissimilarity of observers' prototype from an ideal candidate (as established by the O*NET) and accuracy, after controlling for interview condition, but no support was found for the proposed interaction effects. Possible reasons for theses results are explored and discussed.

Elisa George
Psychology Department
Colorado State University
Fort Collins, CO 80423
Spring 2006

ACKNOWLEDGEMENTS

I would like to acknowledge the contributions and support of my advisor, committee members, and my husband. The continual and valuable guidance offered by my advisor, Eric Heggestad, from proposal through defense, resulted in a greatly improved study and dissertation. Thank you for always finding the time for me. The suggestions and comments of my committee members, Michael Gross, Jack Hautaluoma, and Randy Swaim, also improved the quality of this study. Finally, I am indebted to my husband, Peter, for his unwavering moral support throughout. With much gratitude to all.

TABLE OF CONTENTS

Introduction

In their highly influential meta-analysis, Hunter and Hunter (1984) reported a corrected validity coefficient between the employment interview and supervisory ratings of .14 for entry-level jobs. In comparison to the validity coefficients of alternative predictors of performance such as an ability composite (.53), work samples (.54), and biographical inventories (.37) (Hunter & Hunter, 1984), the employment interview did not fare well. Despite the weak meta-analytical results, however, the employment interview has long been, and continues to be, one of the most commonly used selection devices within organizations (Arvey & Campion, 1982; Wagner, 1949; Wilk & Cappelli, 2003), which led Wagner to state:

> Even though the interview were thoroughly repudiated, it probably would not be abandoned; there seems to be a certain human curiosity which can be satisfied in no other way than by "seeing the man in the flesh." (1949, p. 42)

Given the commitment to the employment interview by practitioners, industrial-organizational psychology has searched for ways to improve its reliability and validity. In particular, researchers have called for greater standardization of the interview, emphasizing consistency in question content and the use of interview rating scales. This standardization later was referred to as "structure." For example, in their review article of structure in the selection interview, Campion, Palmer, and Campion defined structure as "any enhancement of the interview that is intended to increase psychometric properties

by increasing standardization or otherwise assisting the interviewer in determining what questions to ask or how to evaluate responses" and identified 15 different elements contributing to structure (1997, p. 656). In particular, using the same questions was regarded as "the most basic way to convert the interview from a conversation into a scientific measurement" (1997, p. 663). Since Hunter and Hunter's study, many meta-analyses have demonstrated the superior criterion-validity of the structured interview (Huffcutt & Arthur, 1994; Marchese & Muchinsky, 1993; McDaniel, Whetzel, Schmidt, & Mauer, 1994; Wiesner & Cronshaw, 1988; Wright, Lichtenfels, & Pursell, 1989).

Another line of research has explored the interview from the interviewer perspective, emphasizing the social context of the interview and the variance in interviewer cognitions (Carlson & Mayfield, 1967; Dipboye, Gaugler, Hayes, & Parker, 2001; Dougherty, Ebert, & Calender, 1986; Graves & Karen, 1992; Heneman, Schwab, Huett, & Ford, 1983; Mayfield & Carlson, 1966; Pulakos, Schmitt, Whitney, & Smith, 1996; Rowe, 1963; Valenzi & Andrews, 1973; Van Iddekinge, Sager, Burnfield, & Heffner, 2004; Zedeck, Tziner, & Middlestadt, 1983). This research has shown that interviewers differ markedly in their assessment of interviewees and in the ways in which they use information from the interview. Citing individual differences in interviewer competence, Dreher, Ash, and Hancock questioned the interpretability of much of the then existing interview research due to the practice of collapsing data across interviewers:

> The current domain of validity coefficients is largely based on a flawed methodology. Studying the "interview" and not taking individual differences in interviewer competence and rating style into account can result in biased estimates regarding the usefulness of this approach to selection. (1988, p. 323)

As a result of the recognition of interviewer differences, several researchers have called for further investigation of the variables which may influence interviewer accuracy and

validity (Graves & Karren, 1999; Van Iddekinge, 2004). The purpose of this research was to examine the influence of two individual difference variables, general intelligence and the prototype of the ideal candidate held by the interviewer, on interviewer accuracy under various degrees of interview structure.

Criterion Validity of the Employment Interview

Several meta-analysis followed Hunter and Hunter (1984) which have indicated the superior criterion-validity of the structured interview, with corrected validity coefficients ranging from .39 to .62. For example, the following meta-analyses reported the following corrected validity coefficients with regard to supervisory ratings: Wiesner and Cronshaw (1988) reported .62 for structured interview across a range of work-related criteria (corrected for range restriction and criterion unreliability); Wright et al. (1989) reported .39 with respect to a group of situational and past-behavioral structured interviews (corrected for unreliability of the predictor and the criterion, but not range restriction); Marchese and Muchinsky (1993) reported a mean validity coefficient for the structured interview of .45 with respect to job performance criteria (corrected for unreliability of the interview and criteria and range restriction); and McDaniel et al. (1994) reported a mean validity coefficient of .44 for the structured interview (corrected for range restriction and criterion unreliability)[1].

In their meta-analysis, Huffcutt and Arthur (1994) examined the effects of degrees of structure on interview validity and provided a taxonomy of structure which has now become standard[2]. Structure was conceptualized along two dimensions: the nature of the question format and response scoring. Interview questions were placed on a continuum,

ranging from the absence of formal constraints, such as the unstructured interview (Level 1); limited constraints, such as the stipulation of topical areas (Level 2), represented by Ghiselli (1966); the pre-specification of possible questions from which interviewers could choose and probe responses (Level 3), represented by Janz (1982); and complete standardization in which all applicants are asked the same questions and response probes are prohibited (Level 4), represented by Latham, Saari, Pursell, and Campion (1980). Response scoring was also conceived along a continuum of structure ranging from a single overall evaluation of the interview (Level 1), represented by Ghiselli (1966); multiple evaluations along pre-established dimensions or traits (Level 2), represented by Janz (1982); and evaluation of responses to individual questions with a pre-established benchmark answer (Level 3), represented by Latham et al. (1980). As shown in Table 1, based on their location along these continuums, interviews may be categorized into four levels of structure.

Using this taxonomy in their meta-analysis, Huffcutt and Arthur (1994) re-analyzed the studies used by Hunter and Hunter (1984) and found support for a relationship between increased validity and increased structure, although the difference between Structure Level 3 and Structure Level 4 was minimal. Validity coefficients for each level were .20 (Structure Level 1), .35 (Structure Level 2), .56 (Structure Level 3), and .57 (Structure Level 4) (corrected for criterion unreliability and range restriction). Consistent with Huffcutt and Arthur's taxonomy, interview research has tended to focus on the following levels of structure in the interview: the unstructured interview (Level 1), the behavioral interview (Levels 2 and 3), and the situational interview (Level 4).

Table 1.

Levels of Structure in the Interview.

		Interview Question Standardization			
		Level 1	**Level 2**	**Level 3**	**Level 4**
Response Scoring Standardization	**Level 1**	Structure 1	Structure II	Structure II	Structure III
	Level 2	Structure II	Structure II	Structure III	Structure III
	Level 3	Structure II	Structure III	Structure III	Structure IV

Note. Adapted from Huffcutt and Arthur (1994). "Level 1 question standardization was no constraints, Level 2 was limited constraints, typically on the topical areas; Level 3 was precise specification of the questions form which interviewers could choose or follow-up; Level 4 was asking the exact same questions with no choice or follow-up. Level 1 response scoring was a global assessment; Level 2 response scoring was assessment along multiple established criteria; Level 3 was evaluation of each individual response according to reestablished answers." (p. 187)

Frequency of Use

It is generally thought that interviewers resist the use of the structured interview (Dipboye, 1997; Terpstra & Rozell, 1997; Van der Zee, Bakker, & Bakker, 2002). Lievens and De Paepe (2004) used Huffcutt and Arthur's taxonomy in a study of the interviewing practices of Flemish Human Resource (HR) professionals. As there are few laws in Belgium that regulate selection practices, HR professionals have a great deal of freedom in structuring their interviews. These researchers found that the highest percentage of usage (29%) occurred with Structure Level II (pre-determined topics and ratings across multiple dimensions). Respondents indicated that they utilized the lowest level of interview structure (no constraints on the type of questions and a global evaluation) 14% of the time. Only 3% of the respondents indicated the use of Structure Level III (pre-determined questions and probes, among which interviewers may choose, and an evaluation of responses to each question). None of the respondents indicated the use of Structure Level IV interviews (using the exact same questions, with no choice or follow-up, and evaluation of each response).

The Unstructured Interview

The unstructured interview consists of a free format, resembling a conversation (Campion, Palmer, & Campion, 1997) and has displayed lower validities, as well as lower reliabilities, than more structured interviews. Schmidt and Zimmerman (2004) have argued that the interrater reliabilities reported in a meta-analysis performed by Conway, Jako, and Goodman (1995) are the best estimates of interview reliability. Reliability was estimated by the correlation of the interview ratings of two independent raters who

interviewed the same interviewees on two different occasions. No corrections for artifacts were made. Conway et al. reported a reliability of .37 for the unstructured interview. Schmidt and Zimmerman also have argued that the best estimate of the validity of the interview consists of a subset of those reported in the meta-analysis by McDaniel et al. (1994) in which supervisory ratings were gathered for research (as opposed to administrative) purposes. McDaniel et al. reported a corrected validity coefficient of .38 for the unstructured interview. In addition, Wiesner and Crenshaw (1988) reported a corrected mean validity for the unstructured interview of .31 and Huffcutt and Arthur (1996) reported a corrected mean validity for the unstructured interview of .20. Based on the foregoing, the inter-rater reliability of the unstructured interview is probably around .37 and the criterion-related validity is probably between .20 and .38.

The Patterned Behavioral Description Interview

Although Janz (1982; 1989; Janz, Hellervick, & Gilmore, 1986) is usually cited in discussions of the patterned behavior description interview (PBDI), precursors may be found in Hovland and Wonderlic (1939), McMurry (1947), and Ghiselli (1966). The underlying premise of this form of interview is that the best predictor of future behavior is past behavior, sometimes referred to as the behavioral consistency principle (Janz et al., 1986).

As early as 1939, Hovland and Wonderlic (1939) developed the Diagnostic Interviewer's Guide (D.I.G.), in which topical areas of work history, family history, social history, and personal history were to be consistently discussed during the interview. A standardized scoring guide was included and interviewer training was

described. The use of the D.I.G. resulted in improved interview reliabilities and significantly higher scores for applicants remaining on the job for 12 months than dismissed applicants.

McMurry described the patterned interview as a "printed form containing specific items to be covered and providing a uniform method of recording information and rating the interviewees' qualifications" (1947, p. 263). The patterned interview was based on a job specification; designed to measure character traits, such as stability; clinical in approach; and dependent on the competence and training of the interviewer.

Ghiselli (1966) recounted his personal experience over a 17-year period with a standardized interview of candidates for the job of stock broker. All applicants were asked questions regarding personal data, such as their age and marital status, and four questions regarding previous college, military, and occupational experience[3]: What was done? Why was it done? What activities were performed? How well had the applicant done? The interview concluded with a question regarding why the applicant wanted to become a stock broker. Ghiselli reported a validity coefficient with success on the job of .51 (corrected for range restriction).

Probably the best known advocate for the PBDI has been Janz (1982; 1989; Janz et al., 1986). To the behavioral consistency principle, Janz added two corollaries: The more recent the past behavior, and the more long-standing the past behavior, the greater the predictive power of the behavior. The development of behavioral description questions should begin with a job analysis, using the critical incident technique (Flanagan, 1954), in which incidents of particularly effective or ineffective behavior are gathered. Incidents were to be sorted into performance dimensions, usually five to ten.

Although many dimensions combine maximal and typical performance, if a dimension were laden with maximal performance, e.g., technical skills and knowledge, Janz recommended using ability tests, rather than the PBDI format, as he believed the PBDI was best suited to assessing typical performance. "[M]aximum performance focuses on competencies, whereas typical performance focuses on choices" (Janz, 1989, p. 164).

The dimensions and critical incidents were to be used to develop questions which consisted of writing question stems, "which locate a particular instance from the applicant's past and focus the applicant on that type of event or environment" and probes, which "seek out exactly how the applicant behaved and what the consequences of the behavior were" (1986, p. 64). In addition, Janz et al. recommended the use of superlatives in creating question stems - most, last, least - as the use of the superlative would elicit specific events in the mind of the interviewee and would enable the interviewer to place the incident among similar incidents.

With respect to the job of regional representative and the performance dimension "establishing new client contacts," an example of a question stem was: "Tell us about the most difficult new client contact you made in the last six months," and the recommended probes were: "What was the obstacle you faced?" and "What did you say when you were stumped?" (1986, p. 64). Similar questions and probes were to be developed for each dimension. Although one could conduct the interview around the job performance dimensions, Janz and colleagues recommended organizing "the performance topics into groups based on major pattern divisions of the applicant's experience, e.g., recent work experience, job-related work experience, educational experiences, and job-related interpersonal experiences" (p. 67). Although every interview would include questions

related to each dimension, not every question of the pattern would be asked of every applicant.

Janz et al. (1986) strongly recommended either recording the interview or note-taking. At the conclusion of the interview, or after listening to the recorded interview at a later time, applicants were to be assessed on each dimension and placed in 20% percentiles, e.g., bottom 20%, middle 20%, top 20%. If more than one interviewer was involved, it was recommended that their thoughts be compared. If the dimensions differed in importance, weights could be assigned to dimensions and a weighted sum calculated. The applicants with the highest scores were to be made offers.

There have been several individual studies on the reliability and criterion-validity of the PBDI, but I will focus on the meta-analytical results. In their meta-analysis of reliabilities, Conway et al. (1995) identified five levels of structure, from 1 (no formal constraints) to 5 (primary questions specified and no follow-up probing), with Levels 3 (pool of primary questions provided and interviewer choice allowed) and Level 4 (primary questions specified and follow-up probing allowed) perhaps the closest approximations to the PBDI method. Interrater reliabilities were reported as .56 for Level 3 and .66 for Level 4. Taylor and Small (2002) reported a mean reliability for the behavioral interview of .77, when anchored rating scales were used, and .73, when anchored rating scales were not used. In a meta-analysis comparing the behavioral interview to the situational interview, Taylor and Small (2002) reported corrected validity coefficients with performance ratings for the behavioral interview without anchored rating scales of .47, and with anchored rating scales of .63 (corrected for criterion unreliability and range restriction), although the number of studies on which these results

were based was not large. Based on the foregoing, the inter-rater reliability of the PBDI is probably between .56 and .77 and the criterion-related validity is probably between .47 and .63.

The Situational Interview

Latham (1989) developed the situational interview (SI) and he and his colleagues remain ardent supporters of its validity and usefulness (Latham, 1989; Latham et al., 1980; Latham & Sue-Chan, 1999; Mauer, Sue-Chan, & Latham, 1999). Latham (1989) has described the situational interview as a combination of critical incident and goal setting techniques. "The purpose of the situational interview is to identify a potential employee's intentions by presenting that person with a series of job-related incidents, and asking what he or she would do in that situation" (1989, p. 171).

As with the PBDI, questions should be developed from a thorough job analysis, based on critical incidents. In contrast to the PBDI, however, it is not past behavior that is explored, but future behavior regarding what the interviewee would do in certain situations. Proponents of the SI have emphasized, however, that future-looking questions are not sufficient for the SI. The question must also present the interviewee with a dilemma, i.e., mutually exclusive courses of action, the purpose of which is to minimize socially desirable responding (Latham & Sue-Chan, 1999; Mauer et al., 1999). In addition, the SI differs from the PBDI in that each question is scored, using a behavioral anchored scoring guide, which singularly reflects a particular organization's setting (Mauer et al., 1999).

The following is an example of a situational question and scoring guide:

> Your spouse and two teenage children are sick in bed with a cold. There are no relatives or friends available to look in on them. Your shift starts in 3 hours. What would you do in this situation?

Answers were to be evaluated based on the following scoring guide, with *5* indicative of the answer for a best performer and *1* indicative of the answer for a very poor performer:

> (1) I'd stay home – my spouse and family come first.
> (3) I'd phone my supervisor and explain my situation.
> (5) Since they only have colds, I'd come to work. (Latham et al., 1980, p. 424)

In their review of the SI, Mauer et al. (1999) reported inter-rater reliabilities ranging from .76 to .87. Taylor and Small (2002) reported a mean reliability for the situational interview of .79. In their meta-analysis, McDaniel et al. (1994) conducted a separate analysis of the SI and reported a corrected mean validity with job performance of .50. Structure Level 4 of Huffcutt and Arthur (1996) consisted of the same degree of structure as the SI and they reported a corrected validity coefficient of .57. In their meta-analysis of the situational interview, Latham and Sue-Chan (1999) reported a corrected mean validity coefficient of .47 (corrected for criterion unreliability and range restriction), based on what they considered to be the best set of underlying studies. Taylor and Small (2002) reported a corrected mean validity coefficient for the situation interview, including anchored rating scales, of .47, although they did not limit the SI studies to those which presented a dilemma. Based on the foregoing, the inter-rater reliability of the SI probably lies between .76 and .87 and the criterion-related validity probably lies between .47 and .57.

In general, the foregoing meta-analyses support the superior inter-rater reliability and criterion-related validity of the PBDI and the SI compared to the unstructured interview, although fewer meta-analyses have been performed for the PBDI. Despite the

similarity in terms of inter-rater reliability and criterion-related validity, however, comparative research on the two techniques indicates that the two methods may be quite different in other regards.

Comparison of the PBDI and SI

Some of the discussion regarding the two structured interviews has focused on their theoretical foundations. Given that the SI is based on assessing applicant intentions, citing goal setting theory (Locke & Latham, 1984), Latham (1989) has argued that the SI is "among the very few, if not the only, interview technique grounded in theory" (p. 171). Motowidlo (1999), however, has argued against the idea that there is a fundamental theoretical difference between the two interview forms. If past behavior is "presumed to reflect choices, and not mindless, reflex actions. . . . [those choices] are shaped by goals and intentions" (p. 181). In addition, if answers to SI questions are predictive of job performance, it is because intentions are stable and the intended behavior is under our volitional control. But if intentions are stable and behavior is volitional (ignoring the possibility of impression management), then "intended behavior will closely resemble the actual behavior in all past and future occurrences of the specified situation" and there is no clear theoretical distinction between the two types of interview[4].

Some research, however, does raise questions regarding the empirical similarity between the two techniques. Pulakos and Schmitt (1995) compared the PBDI and the SI under controlled and similar conditions. Parallel experience-based and situational questions were written for a set of seven job dimensions, based on the job analysis for a highly complex job in a federal agency (planning and organizing, relating to others,

evaluating information and decision-making, initiative and motivation, adaptability, meeting physical requirements, and demonstrating integrity). Due to the similarity in question content, the same behaviorally-anchored rating sale was used for each format. Incumbent interviewees were randomly assigned to either the situational or experience-based interview, interviews lasted approximately an hour, and probing was allowed. Interviews were conducted by panels of three. At the conclusion of the interview, each interviewer formed independent ratings on the dimensions and the panel was asked to reach consensus, if differences existed greater than one point. Although the means, standard deviations, inter-rater reliabilities, and factor analysis indicated little difference between the two interview formats, only the PBDI showed a significant relationship with supervisory ratings of performance ($r = .32$), while the SI did not ($r = -.02$), and the difference between the two validity coefficients was significant. These researchers speculated that the results may have been the result of not evaluating each SI question, the more "superficial" nature of some of the responses under the SI condition, or the complex nature of the job in comparison to previous studies using the SI format.

Huffcutt et al. (2001) conducted two replication studies of Pulakos and Schmitt (1995), with officer training candidates for the Canadian military and incumbent district managers for a national merchandise chain. In both studies, BDI and SI questions were developed for each dimension related to the job[5]. (For the Canadian officer candidates, the dimensions assessed by both interviews types were: emergent leadership, decision-making, planning and organizing, working with deadlines, addressing poor performance, and resolving group conflict; for the district managers, the dimensions were: initiative, organizing & planning, handling crisis, innovation, motivating others, persuading others,

motivating substandard performers, consideration and support, conflict resolution, and managing peer conflict). However, in contrast to the Pulakos and Schmitt design, the BDI and SI questions were asked within the same interview, with the SI questions asked first in approximately half the interviews and the BDI questions asked first in the other half of the interviews. Interviewers were allowed to probe following the BDI questions in the officer training study, but not in the district manager study. Also in contrast to the Pulakos and Schmitt study, 5-point anchored rating scales were developed for each question and answers were immediately evaluated. Similar to the Pulakos and Schmitt results, in both studies, the BDI was significantly related to performance ratings or rankings ($r = .47$ for the officer trainees and $r = .31$ for the district managers), but the SI was not ($r = .20$ for the officer trainees and $r = .02$ for the district managers) and the difference between the validity coefficients was also significant.

Huffcutt et al. (2001) also performed a multi-trait multi-method (MTMM) analysis. In both studies the mean correlation between SI and BDI questions rating the same characteristics was either lower or not very different (.09 in the officer trainee study and .05 in the district manager study) than the mean correlation between SI and BDI questions ratings different characteristics (.10 in the officer trainee study and .03 in the district manager study). Based on their results, Huffcutt et al. (2001) concluded that the BDI was the more effective predictor for higher-level positions. In addition, in light of the MTMM analysis, they suggested that the SI and BDI assess different constructs – at least for higher-level positions

In contrast to the above studies, Latham and Skarlicki (1995) compared the SI to the PBDI in assessing organizational citizenship behaviors toward the organization

(OCBO) and toward the individual (OCBI) in a group of university faculty members. A behaviorally-anchored rating scale was developed for the SI, while the rating scale used to assess each PBDI question did not include behavioral anchors. Although the coefficients alphas for each OCB dimension were similar between the SI and the PBDI, the inter-rater agreement for each OCB dimension was higher for the SI questions. In addition, the SI questions correlated significantly with OCBO and OCBI, while the PDBI questions did not correlate significantly with either dimension. Latham and Skarlicki attributed the greater effectiveness of the SI questions, in part, to the higher inter-rater agreement resulting from the use of the behaviorally-anchored scoring guide and the absence of probes in the SI format.

In a comparative study between a general-question interview, an SI, and PBDI, Conway and Peneno (1999) investigated the construct validity of the three techniques. Two rounds of interviews were conducted with applicants for the position of resident assistant in a college resident hall. After meeting minimum qualifications, applicants were asked eight general questions in the first interview, the focus of which was the applicant's motivation for wanting the job and their expectations and basic knowledge of the job. A critical incident job analysis was developed and eight job dimensions were identified, of which five were assessed with an SI and a PBDI question (being a role model, programming, helping skills, staff relations, and community development). In order to cloak the job dimensions, the SI questions were asked first. For the first round and the second round interviews, responses to each question were rated on an anchored five-point scale. Ratings were made independently. Applicants were also assessed on the big five personality traits, cognitive ability, and prior leadership experience.

Although SI and PBDI questions were significantly and highly correlated (r = .85) and the general interview was significantly correlated with the SI (r = .38) and the PBDI (r = .31), each interview technique displayed somewhat different correlations with the assessed variables. The general interview was significantly correlated with applicant Agreeableness (r = .17) and Neuroticism (r = .16), but none of the Big Five personality traits were significantly correlated with the SI and PBDI. Although both the SI and the PBDI were significantly correlated with applicant leadership experience, the PBDI was more strongly correlated with prior leadership experience (r = .43) than the SI (r = .29). In addition, none of the interviews was significantly correlated with applicant cognitive ability. The SI and PBDI displayed convergent validity under a MTMM analysis for all five dimensions (mean r = .50), but there was no evidence of discriminant validity, leading these researchers to conclude that there was no evidence that the questions were assessing the intended construct. Conway and Peneno (1999) speculated that the SI and PBDI did share a large proportion of job-related variance, e.g., perhaps tacit knowledge, if not cognitive ability, but also that the SI questions tended to indicate an understanding of the job, while PDBI questions were more indicative of past experience.

In summary, the meta-analytical evidence suggests that the use of the PBDI or the SI results in similar increases in inter-rater reliabilities and criterion-related validities over the unstructured interview, although the SI may result in slightly better inter-rater reliabilities (.76 to .87) than the PBDI (.56 to .77) and the PBDI may result in slightly higher criterion-related validities (.47 to .63) than the SI (.47 to .57). Motowidlo (1999) has argued that the theoretical difference between the SI and PBDI cannot be traced to differences in the role of intentions. However, the comparative research on the SI and

PBDI formats has resulted in mixed evidence regarding their convergent validity, a lack of evidence regarding their discriminant validity, and evidence that all three interview techniques correlate differently with other variables. As will be explored further in the next section, perhaps the source of the perplexing results resides not in the nature of the interview, but in differences in interviewer perceptions under each of the interview formats.

Interviewer Differences

A series of "paper-people," policy capturing, and validity studies, utilizing both incumbents and applicants, as well as a range of structure in the interview, provide support for differences in how interviewers use information. Interviewers differ in what they regard to be unfavorable facts about applicants (Carlson & Mayfield, 1967; Mayfield & Carlson, 1966), the ways in which they weigh and combine information (Graves & Karren, 1992; Zedeck, Tziner & Middlestadt, 1983), their self-awareness of the ways in which they use information (Graves & Karren, 1992), and their ability to predict performance from the interview (Dipboye, Gaugler, Hayes, & Parker, 2001; Van Iddekinge et al., 2004; Zedeck, Tziner, & Middlestat, 1983).

Mayfield and Carlson (1966; Carlson & Mayfield, 1967) presented a list of items to life insurance managers, who were asked to indicate the favorability or unfavorability of the item with respect to the position of life insurance agent. Although there was a high degree of inter-rater agreement among the managers regarding some items (e.g., earned all his college expenses) there was extreme disagreement among other items (e.g., the applicant is presently active in eight outside groups). When a group of managers were

presented with hypothetical applicants consisting of only items about which there had been disagreement regarding favorability, the inter-rater agreement was quite low, with some of the hypothetical applicants being ranked best and others ranked worst. These researchers concluded that every interviewer has a stereotype comprised of two parts: one consisting of favorable items to which there is a high degree of agreement and a specific stereotype which is different for different interviewers.

Zedeck, Tziner, and Middlestadt (1983) examined the interview ratings of female officers of female candidates for officer training school on nine dimensions (expression, decision-making, self-confidence, performance under stress, persistence, analytical ability, interpersonal sensitivity, self-understanding, and openness) and an overall rating (the degree of interview structure was not specified). These researchers found evidence for individual interviewer strategies and differential weighting of dimensions, with the greatest proportion of interviewer variance explained by two or three dimensions. Virtually no correlation was found between overall interview evaluations and training performance at 6 and 12 weeks (although range restriction may have played a role in these results) and support was found for the greater predictive accuracy of several interviewers.

Graves and Karen (1992) asked 29 professionals from a financial services company to read descriptions of hypothetical candidates which varied on five selection criteria (work experience, education, oral communication, motivation, interpersonal skills) related to customer service positions and to indicate how qualified each candidate was for the job and whether they would recommend hiring the applicant. Participants were also asked to rank order the selection criteria in terms of importance. These

researchers found substantial differences in interviewers' decision processes, including 13 different ways of weighting factors related to the candidate's qualifications and 6 different ways of weighting factors related to the hiring recommendation. Differences in the correspondence between interviewers' professed importance rankings and their actual use of the criteria was also found, with effective interviewers showing a higher degree of correspondence.

In an investigation of the validity of the panel interview, Dipboye, Gaugler, Hayes, and Parker (2001) used an unstructured interview to assess nine dimensions related to an entry-level position in a corrections institution (enthusiasm, cooperation in the interview, self pride, and job stability, as well as the abilities to work with others, follow rules and regulations, accept supervision, perform based on physical condition, and to supervise inmates). These researchers found considerable variability in validity across panels and individual interviewers and concluded that the typical low validity associated with the unstructured interview may be unwarranted: "[O]ur most valid individual interviewer approached the level of validity usually only seen for structured interviews" (2001, p. 47).

In one of the largest investigations of interviewer validity differences Van Iddekinge et al. (2004) investigated the criterion-validity of 64 interviewers (non-commissioned officers or NCOs) of 944 junior NCOs. The interview consisted of a structured past-behavioral interview which assessed six job-related dimensions (adaptability, initiative, integrity, relating to others, leadership, and self-management), using a behaviorally-anchored rating scale. A wide range of criterion-validities were found across individuals in predicting three different performance criteria, ranging from

-.66 to .66 for current performance ratings, from -.76 to .80 for expected future performance rating, and from -.26 to .66 for promotion points. These results could be attributed to sampling error in only the case of the promotion point scores. Van Iddekinge et al. examined several "convenience" variables (gender, race, job experience, and interview attitudes) as correlates of interviewer validities, and found a moderate effect for race, in the case of current performance, and a moderate effect for interview attitudes, in the case of future performance, but encouraged research of more theoretically-based individual difference variables, such as cognitive ability, among others. These researchers recommended that "future research should investigate how different elements of structure affect the psychometric quality and outcomes of interview ratings" (p. 25).

Models of Interviewer Differences. Based on the evidence discussed above, as well as social cognitive theory, several theorists have emphasized the interactive nature of the interview and the importance of the cognitions of the interviewer. Relying on an interactionist perspective, Eder and Buckley (1988) encouraged a perspective which "places a more equal emphasis on dimensions of the person and dimensions of the situation as they interact within the cognitive awareness of the interviewer" (p. 80). Of interest to this research, they discussed the relevance of cognitive schemas to the selective attention, categorization, and combination of information gathered in the interview. Motowidlo (1986) presented a general model of information processing in personnel decisions in which "evaluative judgment is determined in part by the combination of evaluatively positive and evaluatively negative bits of information..." (p. 3). In particular, he argued that interviewer positive and negative judgments would be guided by the interviewer's prototype of the ideal candidate. Dipboye (1992) presented a

multi-stage model of the interview – pre-interview impressions, the interview itself, and post-interview evaluation and decision – with cognitive processes and knowledge structures, such as stereotypes, schemas, and implicit theories, influencing the interviewer in each stage.

In perhaps the most comprehensive model of individual differences which may impact interviewer judgments, Graves (1993) offered a model in which five broad individual difference variables were thought to influence interviewer information processing, which in turn would influence interview judgment and effectiveness: Interview background variables (experience, demographic characteristics, cognitive abilities, personality characteristics), cognitive structure (prototype of the ideal applicant, structure for evaluating applicants, attributional rules, interviewer scripts), interview context, conduct of interview (content and structure, interviewer behavior), and affect (undifferentiated affect, differentiated affect).

For this research, I chose to focus on two of the Graves individual difference variables which could influence interviewer information processes and judgments, one of which is regarded as a stable trait, general intelligence, and one of which is perhaps regarded as more mutable, interviewer prototypes.

Interviewer differences and intelligence. General intelligence has long been associated with the ability to judge others. Vernon (1933) associated intelligence with the ability to judge personality and, in particular, with the ability to judge intelligence in others. Based on the then extant research, Taft concluded that there was a positive relationship between intelligence and judging others analytically – i.e., judgments in which the judge is "required to conceptualize, and often to quantify, specific

characteristics of the subject in terms of a frame of reference," e.g., rating traits (1955, p. 1). Indeed, to the extent that judging others is a specific instance of a general cognitive task, it would be expected that general intelligence would be related to judgmental accuracy (Smither & Reilly, 1987).

Borman (1979a) presented a group of students with videotapes of a recruiting interview and a manager in a problem-solving session, in which the performance on various dimensions was pre-determined (for the recruiter position, the dimensions were creating a positive image for the company, being organized, providing information, asking relevant questions, answering questions, establishing report; for the manager position, the dimensions were: controlling the interview, establishing rapport, reacting to stress, obtaining information, resolving conflict, developing subordinates, and motivating subordinates). True scores were established by the mean ratings on the dimensions by a group of experts. Among other characteristics, accurate perception of performance was correlated with the students' intelligence, high grades, and investigative interests and verbal reasoning.

Smither and Reilly (1987) presented undergraduates with videotapes of five actresses playing the role of customer sales representatives for a telecommunications company, in which the intercorrelations among three dimensions of performance were pre-determined (sales of products and services, authorizing and processing paperwork for repairs, and resolving customer account problems on a computer). Using Cronbach's (1955) rater accuracy scores (elevation, differential elevation, stereotype elevation, and differential accuracy), Smither and Reilly found a positive relationship between intelligence and overall accuracy, stereotype accuracy, and differential elevation.

However, these relationships were qualified in that the effect of rater ability on accuracy was greater when the rating task was moderately difficult, i.e., when the intercorrelations among dimensions was moderate or high. When job dimensions were uncorrelated, intelligence was unrelated to rater accuracy, suggesting to these researchers that the extreme difficulty of the task resulted in lower accuracy for all, irrespective of intelligence. In addition, elevation and differential accuracy were unrelated to rater intelligence.

In an examination of the interview and applicant-organization fit, Jako (1990) found that more intelligent interviewers were better able to identify the dominant personality trait in candidates, but interviewer intelligence was not related to assessing organizational traits or assessing applicant-organization fit. However, Jako also thought that it would be premature to conclude that intelligence is not related to interviewer accuracy in assessing applicant-organization fit, as the task of assessing organizational needs may have been too simple in his study.

Interviewer Differences and Prototypes. Based on the individual differences in the ability to judge others, several researches and theorists have looked to the schemas, stereotypes, prototypes, or implicit theories held by interviewers in the way of explanation. Although many researchers use the term prototype and schema interchangeably, some theorists distinguish between them. For example, Fiske and Taylor (1991) defined a schema as:

> [A] cognitive structure that represents knowledge about a concept or type of stimulus, including its attributes and the relations among those attributes. . . . As people's theories and concepts about the world, schemas are concerned with the general case, abstract generic knowledge that holds across many particular instances. (1991, p. 98)

In contrast, they defined a prototype as:

> [A] most typical or prototypical instance best representing the category. The prototype is the "central tendency" or average of the category members. (1991, p. 106)

In addition, prototypes resemble instances in that "all their known attributes are filled in, even if all the attributes are not directly relevant to the category membership," while some features may be ignored in schemas (1991, p. 117). In a similar fashion, in his discussion of the role of prototypes in the interview process, Guion also distinguished between a stereotype and a prototype, defining a prototype as "a carefully and systematically developed ideal to be achieved. . . . A set of attributes that not only describe the desired candidates but distinguish them from those less desired" (1987, p. 202).

While theorists differ in their use of the terms schema, stereotype, and prototype to describe the relevant cognitive structure, most agree that interviewers hold a prototype of the ideal candidate and that this prototype influences their cognitive processes to some extent. Information from the interview is attended to, categorized, and combined on the basis of the schema or prototype, and such pre-existing prototypes may lead to biases or distortions in the interviewer's perceptions of the characteristics of the interviewee (Dibpoye, 1992; Eder & Buckley, 1988).

Rowe (1984) thought that selection decision-makers hold a cognitive structure for "good workers," which although based on a job analysis, is unique and subject to variation in terms of the dimensions included, the relevance of the dimensions, and the weight placed on different dimensions for each interviewer. In addition, decision-makers hold cognitions of the prototypical "ideal worker," abstracted from the good worker

prototype, and applicants are accepted as good workers based on the match between the dimensions of the prototype and the dimensions of the applicant, as perceived by the decision-maker.

Motowidlo (1986) argued that the interviewer prototype determines which pieces of interviewee information are viewed positively and negatively, with characteristics similar to the prototype being viewed positively, and those opposite of the prototype, being view negatively. Consequently, "interviewers with an accurate prototype will evaluate cues correctly; that is, they will evaluate them just as they are signed (positively or negatively) in the true score domain" (1986, p. 13).

Dalessio and Imada asked five interviewers to evaluate applicants for the position of telephone operator on various dimensions and on their overall desirability under a structured board interview which "elicited verbal responses and role-playing simulations" (1984, p. 70). Several weeks later, the interviewers rated their ideal candidate and themselves on the same rating instrument. These researchers found that similarity between the rating of the applicant and the ideal candidate was positively related to a favorable overall decision and that similarity to an ideal candidate demonstrated a stronger relationship with a favorable decision than similarity with self.

During the 1980s, a debate occurred in the literature regarding the degree to which raters accurately perceive behaviors related to performance versus the influence of prototypes on performance ratings. Borman (1978, 1979a, 1979b) was considered to be representative of the traditional accuracy model in which performance evaluation was regarded as a three-step process involving: "(a) observing worker-related behavior, (b) evaluating each of these behaviors, and (c) weighting these evaluations to arrive at a

single rating on a performance dimension" (1978, p. 141). In contrast, relying on cognitive categorization models, Lord and others (Lord, 1985; Nathan & Lord, 1983; Padgett & Ilgen, 1989) proposed that raters engage in a "heuristic process in which information is automatically stored as part of a prototype-based category. . . . [which] serves as the basis for subsequent behavioral ratings rather than the originally observed behavior" (Nathan & Lord, 1983, p. 103). Several studies attempted to determine which model was more accurate and found support for both processes (Nathan & Lord, 1983; Padgett & Ilgen, 1989):

> "[S]upport for the traditional model of performance rating implies that dimension-based encoding of information can occur. Thus, at least under our optimal conditions, ratings contain significant true score variance. Second . . . naturally occurring categories will be used to simplify information processing and thereby make the world more orderly than it already is . . . . the correlation among rated dimensions will almost always be higher than true correlations, due to the nonrandom errors or simplifications of raters." (Lord, 1983, p. 112)

In summary, both social cognitive theory and research indicate that prototypes – i.e., fairly robust, articulated concepts of the ideal candidate – play a role in the assessment of others and that these prototypes vary across observers.

Hypotheses

This study focused on the accuracy of interviewer assessments of dimensions across three interview formats: unstructured, the PBDI, and the SI. Given that this was a controlled experiment, the "unstructured" interview format was considerably more structured than a typical unstructured interview in which responses to questions often prompt the interviewer to ask unanticipated follow-up questions. In this study, the same questions were asked of each of the candidates in the unstructured format, although those

questions were of the sort generally asked in unstructured interviews – questions which focused on the candidate's attitudes, opinions, or preferences, as well as general ways of responding in certain situations (see Appendix B). As such, the unstructured interview in this study is best regarded as an attitudinal interview and will be referred to as an "attitudinal interview" (AI) in what follows.

Participants were randomly assigned to one of the three interview formats and were asked to view video recordings of three people being interviewed for the job of graduate teaching assistant under each interview format. (As participants did not actually conduct an interview, but simply watched a video of an interview being conducted, their experience may be compared to a passive member of an interview panel. Given this perspective, participants are also referred to as "interviewers.")

Based on the meta-analysis cited above, there is substantial evidence that structure generally improves the criterion validity of the employment interview. Given that structure improves criterion validity and given that judgmental accuracy may be expected to be a necessary condition for criterion validity, the following hypothesis is proposed:

> *Hypothesis 1*: There will be a positive relationship between interviewer accuracy and the degree of structure of the interview, such that interviewer accuracy will be least in the AI format and greatest in the SI format, with the PBDI format resulting in a degree of accuracy which falls between the attitudinal and SI formats.

The comparative research on the unstructured, SI and PBDI formats suggests that differences exist between the techniques in terms of the constructs assessed and their respective relationships with other variables (Conway & Peneno, 1999; Huffcutt et al., 2001; Latham & Skarlicki, 1995). In addition, there is mixed evidence for the convergent validity of the SI and PBDI and no evidence for their discriminant validity (Conway &

Peneno, 1999; Huffcutt et al., 2001). Although researchers have generally looked to subtle differences in the nature of the interview (e.g., how the rating scale was used) or the context (e.g., the level of job complexity), another source of variance may be differences in interviewer cognitions as a result of the interview format and individual interviewer differences.

The interviewer difference literature has repeatedly found a high degree of variance across interviewers in perceived interviewee characteristics (Carlson & Mayfield, 1967; Mayfield and Carlson, 1966), in how information is weighed and combined (Graves & Karren, 1992; Zedeck et al., 1983) and in interviewer validity (Dipboye et al, 2001; Van Iddekinge et al, 2004, Zedeck et al., 1983). In short, these studies find that some interviewers make better judgments than others. In addition, both theory and research may be used to argue for a relationship between interviewer intelligence and interviewer judgments. One could also argue that a more structured interview format requires less cognitive work on the part of the interviewer. Therefore, the following hypothesis is proposed:

> *Hypothesis 2*: General intelligence will moderate the relationship between interview accuracy and interview structure, such that greater general intelligence will result in greater accuracy under the AI format and will not result in greater accuracy under the SI format, with the relationship between interviewer intelligence and the PBDI format falling in between these two extremes.

Both theory and research may be used to argue for degrees of accuracy in interviewer perceptions and for the influential role of an interviewer's prototypes on the evaluation of interviewees. In addition to individual differences, perhaps the interview context, i.e., the degree of structure, influences the extent to which interviewer prototypes influence interviewer judgments. Prior research has shown that a reliance on cognitive

categories or prototypes may result in less evaluative accuracy (Nathan & Lord, 1983; Padgett & Ilgen, 1989). Given that the SI format is the most structured format, an interviewer's prototype of the ideal candidate may play a lesser role in the interviewer's judgments. Janz also suggested that the PBDI may result in interviewers being "more immune to invalid job stereotypes" than interviewers in an unstructured format (1982, p. 580). In other words, it is proposed that the interviewer's concept of the ideal candidate would influence his or her perceptions of the interviewee. The greater the difference between the interviewer's concept of the ideal candidate and the ideal candidate (as established by O*Net importance ratings), the less their accuracy. And this negative relationship between difference-from-ideal-candidate and accuracy will be greatest in the AI interview format, and somewhat less in the PBDI format, whereas no relationship between difference-from-ideal-candidate and accuracy will be found in the SI format. Therefore, the following hypothesis is proposed:

> *Hypothesis 3:* Interviewer prototypes will moderate the relationship between interviewer accuracy and interview structure, such that the relationship between the interviewer's difference-from-ideal-candidate and accuracy will be negative and greatest in the AI condition and difference-from-ideal-candidate will not be related to accuracy in the SI condition, with the PBDI format falling in between these two extremes.

Method

Participants

Participants were 301 undergraduate students from introductory psychology courses who were given course credit in exchange for participating. The pool consisted of 211 females and 90 males. On average, the students were 19 years old and indicated they had two to three years of work experience and limited prior experience in conducting employment interviews (generally, not having conducted any, or one or two, prior employment interviews). The students self-registered for study sessions and did not know which interview condition would be conducted at particular times. As there were no significant differences between conditions regarding gender-composition, age, work experience, interview experience, or the variables of interest to this study, there is no reason to believe that the self-registration did not approximate random assignment. In total, 99 students participated in the AI condition, 107 in the SI condition, and 95 in the PBDI condition.

Job Analysis

So that I could provide a similar frame-of-reference to both those who participated in developing the study materials and participants, a job analysis was created for a graduate teaching assistant (GTA), based on the Occupational Information Network (O*Net). The O*Net was created and is maintained by the U.S. Department of Labor and

"contains a significant amount of meaningful and reliable job analytic information for a vast array of jobs" (Jeanneret & Strong, 2003, p. 466). For each job within the database, O*Net indicates, in summary and detailed form, the related tasks, job knowledge, skills, abilities (KSAs), work activities, and work context. In addition, O*Net provides an importance value for each KSA, from 0 to 100.

Research has shown that assessors' rating accuracy increases when fewer dimensions are assessed (Gaugler & Thonton, 1989; Thornton, 1992), and given the need to limit the length of the interview in the present study, only a limited number of the KSAs were drawn from the O*Net. I decided to retain those dimensions with importance values greater than 80 which lent themselves to assessment through an interview (e.g., the ability to read and understand information presented in writing carried an importance value of 85, but was excluded because it did not lend itself to assessment through the interview). This approach resulted in selecting seven KSAs, four of which would be assessed by developing interview questions - knowledge of education and training, skill in instructing, skill in critical thinking, and the ability to comprehend oral information – and three of which would be assessed through direct observation - use of English, oral expression, and speech clarity.

Interview Development

Critical Incident Technique. Interview developers (Janz et al., 1986; Latham & Sue Chan, 1996) often recommend basing interview questions on the critical incident technique (Flanagan, 1954). The purpose of the technique is to gather examples of behavior which represent exceptionally effective or ineffective performance with respect

to job-related activities – in this case, the KSAs related to the job of graduate teaching assistant. The incidents should be specific, focus on observable behavior, describe the context, and indicate the consequences (Bernardin & Russell, 1998).

Four current GTAs from Colorado State University (CSU) volunteered to participate in sessions in which critical incidents were developed. Two of the GTAs were math majors and two were computer science majors, with teaching experience ranging from 3 to 5 semesters. In preparing for the critical incident sessions, it became apparent that the O*Net education and training dimension addressed two distinct qualities: 1) Demonstrates knowledge of the principles and methods for designing a curriculum and 2) Demonstrates knowledge of principles and methods for assessing learning. Therefore, this dimension was parsed into these two aspects and the GTAs were asked to develop critical incidents with respect to five KSAs.

In these critical incident sessions, I explained the basis for the study, the genesis of the job analysis, and reviewed the job description and relevant KSAs for which the GTAs would be asked to develop critical incidents. I explained the purpose of the critical incident technique and shared examples of critical incidents for the dimension of knowledge of the principles for designing curriculum and teaching others. For each KSA, I asked the GTAs to recount in writing one example of very effective and very ineffective behavior. In total, 40 critical incidents were developed. An example of a critical incident developed by one of the GTAs with respect to knowledge of the principles for designing a curriculum was:

> On the first day of class, the GTA provides a very detailed syllabus of the material to be covered. The syllabus includes all assignments, due-dates, exams, and describes the method of assigning grades. This is an example of effective behavior because it immediately informs each student of what is expected of

them. This results in students being able to set their own goals for obtaining a certain grade; students could be more organized and prepared for the semester.

These 40 critical incidents served as the basis for question development. In several cases, the nature of the critical incidents offered by the GTAs were very similar, e.g., with respect to measuring learning, several GTAs thought it was important for exams to closely resemble in-class and homework assignments. If the critical incident did not cite behaviors, I tended not to use it, e.g., the GTA assumed a pre-existing knowledge framework with respect to the class. If the critical incident was not generalizable to other disciplines, I also did not use it in developing questions, e.g., the GTA supplied both geometric and algebraic solutions to problems. In addition, the critical incidents related to the dimension of oral comprehension were somewhat contrived and did not lend themselves to assessment through the interview and that dimension was dropped from the study. Another overriding goal was to limit the number of qualities assessed through each interview. Based on the foregoing considerations, the critical incidents were used to develop three interview questions for each of the three interview formats for each of the four remaining KSAs to be assessed with interview questions (two related to knowledge of education and training and one each for skill in instructing and critical thinking), resulting in 36 preliminary interview questions.

Unstructured Interview (AI) Development. As unstructured interviews differ greatly from study to study in terms of the questions asked (Harris, 1999), there is no specific format to follow. However, unstructured interviews tend to elicit credentials, general experience descriptions, and self-perceptions, such as likes and dislikes, strengths and weaknesses, goals and attitudes, and what-would-you-do-if statements (Janz, 1982; Janz et al., 1986). Although the hallmark of the unstructured interview is a free format

and a lack of predetermined questions, given the experimental setting, questions were designed to assess the same dimensions as the PBDI and SI. The AI questions tended to focus on general experiences and attitudes, and follow-up probes were not developed. The following is an example of an AI question, based on the critical incident regarding syllabus development shown above:

What kind of syllabus do you like to provide for your students?

PBDI Development. In developing the PBDI, the recommended procedures of Janz et al. (1986) were followed. Each dimension was evaluated as to whether it reflected typical versus maximal performance. If a dimension reflects maximum performance, e.g., knowledge, one should seek out other possible means of testing for the dimension. Among the dimensions relevant to the job of graduate teaching assistant, the education and training dimensions could perhaps be regarded as reflective of maximum performance. However, Janz et al. also noted that few dimensions reflect purely maximal or typical performance. In addition, what may be most relevant to the job of graduate teaching assistantship would be the way in which one behaviorally instantiates theoretical knowledge of education and training. Thus, it was possible to write questions regarding specific behavioral examples of someone using principles and methods for teaching. For each of the KSAs, based on the set of critical incidents, question stems and possible follow-up probes were developed. Also consistent with Janz et al.'s recommended procedures, the question stems queried about an extreme incident, e.g., a most difficult time, a challenging time, a recent time. For example, the past-behavioral question, along with follow-up questions, based on the critical incident related to syllabus development was:

Tell me about the last syllabus you developed.
What sorts of information did it include?
How long was it?
How was it delivered to your students (e-mail, physical)?

SI Development. Latham and colleagues (Latham, 1989; Latham & Sue-Chan, 1996) have documented the steps to be used in developing an SI interview and these steps were followed in this study. Those steps are (ignoring steps which are relevant only if one is in a position to perform a criterion-related validity study): 1) Conduct a job analysis, using the critical incident technique, 2) Turn relevant critical incidents into a "what would you do if" question, 3) Develop an anchored scoring guide for each question, and 4) Conduct a pilot study in order to eliminate questions for which the answers show no variance or to which interviewers cannot agree on the scoring. Latham and Sue-Chan (1999) have also emphasized, particularly in later writings, that the questions must present the interviewee with a dilemma or mutually exclusive course of action. Therefore, each SI question presented the interviewee with a dilemma. For example, the SI question, based on the critical incident related to syllabus development was:

> Suppose that you are behind in developing your course material for the upcoming semester. Classes begin next week. How would you handle the preparation of your syllabus in this situation?

Pilot Testing. In order to test the content validity of the initial set of questions, three subject matter experts (SMEs; fourth- and fifth-year graduate students in the Industrial-Organizational Psychology program at CSU) were asked to sort each of the 36 questions into one of the four dimensions intended to be assessed through the interview. In general, in order for a question to be considered indicative of a KSA, two of the three SMEs needed to agree on the KSA assessed by the question. Based on this analysis, insufficient agreement existed regarding one dimension (demonstrates knowledge of the

principles and methods for designing curriculum) and this dimension and the related questions were removed from the study.

Three GTAs, who were English majors, with two to five semesters of experience teaching of an introductory composition course at CSU, allowed me to conduct and tape-record mock employment interviews with them, utilizing the 36 questions. These mock interviews revealed that some questions needed to be modified in order to address the nature of a composition class (e.g., exams are not generally given) and questions which resulted in little variance in answers were removed. Based on the sorting exercise and the mock GTA interviews, the two best questions for each interview format related to three remaining KSAs were chosen (knowledge of principles and methods for assessing learning, skill in instructing others how to do something, and critical thinking) and a behaviorally-anchored rating scale related to each question was developed.

Rating Scale Development

As noted above, an important way in which interviews may differ is in the type of rating scale used. The unstructured interview has been associated with an overall rating (Huffcutt & Arthur, 1994). Although Janz et al. (1986) recommended rating applicants on dimensions in 20% percentiles, researchers have used the PBDI technique with behaviorally-anchored rating scales with respect to dimensions (Pulakos & Schmitt, 1995; Van Iddekinge et al., 2004) and behaviorally-anchored ratings scales with respect to questions (Huffcut & Arthur, 1994), and Janz (1989) recommended that research be conducted with benchmark answers. Latham and colleagues (Mauer, Sue Chan, &

Latham, 1999) have insisted that the SI include an interview scoring guide for each question, with behavioral anchors indicating best and worst answers.

Although I considered varying the rating scale as well as the nature of the interview questions in each condition, this approach would have confounded the nature of the interview questions with the nature of the ratings scale, resulting in an inability to distinguish the effects of question-type from the effects of rating-scale type. For this reason, a question-based behaviorally-anchored rating scale was developed and kept constant across interview conditions. For each question developed to assess a KSA, behavioral anchors for the most effective and least effective behavior were developed from the critical incidents. For example, the behavioral anchors related to syllabus development ranged from a weak answer, in which the candidate indicated that he or she rarely relied on or referred to a syllabus, to a strong answer, in which the candidate indicated that he or she provided a highly detailed syllabus, including all assignments, due dates, exams, and grading methods.

In all interview conditions, at the conclusion of each interview, the participants were asked to rate the interviewee's response to each question. With respect to the three KSAs to be directly assessed through observation and not by particular questions (knowledge of English, the ability to communicate through speaking, and the ability to speak clearly), no behavioral anchors were created. The final job analysis is shown as Appendix A and the final interview questions and related rating scales are shown as Appendix B.

Reliability. In order to assess the interrater reliability of the behaviorally-anchored rating scales, $r_{wg(J)}$ values were calculated (James, Demaree, & Wolf, 1984), based on

the participants' ratings. This reliability index measures the degree to which judges agree on a set of judgments (i.e., the ratings) for a single target. Given that there were three GTAs interviewed under three different conditions, nine $r_{wg(J)}$ values were calculated, with values ranging from .79 to .90 (as shown in Table 2).

Table 2.

Interrater agreement for the behaviorally-anchored rating scale.

Interview Condition	GTA 1	GTA 2	GTA 3
AI	.86	.89	.85
SI	.89	.90	.85
PBDI	.86	.79	.86

Scripts and Video Recordings of the Interviews

The tape-recorded mock interviews with the three English majors were transcribed and served as the basis for me to write scripts for actresses to enact during the video-recording of the interviews which would be viewed by the participants. As the tape-recorded mock interviews tended to run an hour or more, it was necessary to abbreviate actual answers. In addition, extraneous material and false starts were removed. Actual answers were also modified if it was regarded as necessary in order to avoid the possible identification of the GTA or any of their students. In order to avoid raters' tendencies toward a halo effect (in which a rater allows his or her ratings on one dimension to influence his or her ratings on other dimensions), the scripts were crafted such that each GTA responded with very effective, moderately effective, or ineffective

answers with respect to the questions related to each of the three KSAs. The targeted levels of each KSA with respect to each GTA are shown in Table 3. In some cases, I exaggerated or modified actual answers in order to reflect the targeted level of the KSA. Despite these modifications, however, in writing the scripts, I attempted to preserve as much of the original answer as possible.

Table 3.

Targeted Levels of GTA Answers

KSA	GTA #1	GTA #2	GTA #3
Knowledge of the principles and methods for assessing learning	High	Low	Moderate
Teaching others how to do something	Low	Moderate	High
Use of logic and reasoning in order to compare and contrast alternative solutions or approaches to problems.	Moderate	High	Low

Three female actresses were hired to play the role of the GTAs and the interviews were video-recorded. Each actress answered the same six questions in the same order (two for each of three dimensions) in each interview condition. Thus, all participants observed the same three actresses in each condition. Follow-up questions were utilized in only the PDBI format and only if the answer content failed to address an important element of the past-situation. Typically, one follow-up question was utilized with respect to four or five question stems in each of the three PBDI interviews.

Despite the fact that tape-recorded answers had been significantly reduced in length, the scripted answers were still too long for most actresses to memorize. As such, the actresses were asked to memorize only the key elements in each response and were allowed to improvise around those key phrases. The resulting attitudinal and situational

interviews tended to be about five to six minutes long, while the past-behavioral interviews tended to be six to seven minutes long.

True Scores

Given that there was no guarantee that crafted answers would be perceived as reflecting the intended level of the KSA, and because the actresses could deviate to some extent from the written scripts, each interview was shown to a group of SMEs in order to establish the latent level of the KSA achieved during video-recording (i.e., the SMEs scores were regarded as true scores). The true-score SME group consisted of four senior graduate students in the Industrial-Organizational Psychology program at CSU, one of which had had recently received her MA degree, while the remaining three graduate students had either defended their dissertations or were in the process of conducting their dissertation research. The SMEs were asked to individually rate each interview answer using the behaviorally-anchored rating scale. Evaluations were then shared and observations discussed until consensus was reached regarding the level of the dimension-related behavior demonstrated by each answer. The SME ratings were in the range of the targeted level of the KSA in 42 of 54 cases (78% of the time). That is, if the intended level of the KSA was high, the SME consensus rating was a six or seven. In the 12 cases (22%) in which the SME ratings was not within range of the intended level of the KSA, 11 of the ratings were within 1 point of the intended level and, in one case, within 2 points of the intended level. Given this degree of correspondence between intended scores and SME scores, the SME scores were treated as true scores for the purposes of determining the accuracy of observer scores.

In the initial phases of this study, I thought that the GTAs employed for the mock and actual interviews would come from various disciples, ranging from liberal arts to the natural sciences. Had GTAs who were not English majors been employed during pilot testing, those GTAs probably would have displayed various levels of knowledge of the English language, the ability to communicate ideas through speaking, and the ability to speak clearly. But given that all the GTAs on which the scripts were based were English majors, and given that all three actresses also were highly articulate and their natural language was English, with few exceptions, the SMEs rated each actress highly on these dimensions. Although the participant ratings showed somewhat more variance, participant ratings were also skewed to the high end of the scale with respect to these qualities. Based on the foregoing considerations, I believe that both the SME and the observer ratings on these dimensions were not highly relevant to the aims of this study and the ratings on these qualities were not included in any of the analyses.

Measures

General Intelligence. The Wonderlic Personnel Test (WPT) was used to assess participant general cognitive ability. It is a strictly-timed, 12-minute intelligence test, consisting of 50 items which include an array of questions from word comparisons to analysis of geometric figures. The distributors of the WPT report test-retest reliabilities of .82 to .94 and representative predictive validities ranging from .27 to .67 (Wonderlic Manual, pp. 20-21, 43). The median and the average WPT score across all conditions was 24, which is the median test score reported by the Wonderlic distributors for first year college students (p. 18).

Participant Prototype. Participants' prototype of the ideal GTA was assessed by asking participants to indicate how important certain KSAs were to their concept of the ideal GTA, using the O*Net's 100-point scale, with *0* indicating the least important and *100* indicating the most important. Given that it is thought that interviewers' prototypes are fairly descriptively rich (Fiske & Taylor, 1991), additional dimensions than those intended to be assessed in the interviews were again drawn from the O*Net, resulting in 25 assessed KSAs. In order to give an additional degree of realism to the ideal candidate survey, participants were given an opportunity to write in and indicate the importance of additional dimensions, although the write-in qualities were not utilized in the study.

KSA Accuracy. Accuracy of participants' ratings was assessed by taking the sum of the absolute differences between the participant's ratings of the candidate on each question and the true score ratings of the candidate on each question, as established by the SMEs. A perfect match between observer ratings and SME ratings would result in a score of 0. Thus, lower difference scores indicate greater accuracy and higher difference scores indicate less accuracy.

Difference-from-Ideal-Candidate. In assessing the influence of participants' prototypes on accuracy, an additional difference score was calculated – the sum of the absolute differences between the participant's importance ratings of the KSAs associated with their ideal GTA and the importance ratings of the KSAs associated with a GTA as established by the O*Net. This difference-from-ideal candidate represents the degree of similarity (or dissimilarity) between the participant's concept of the ideal GTA and the O*Net's evaluation of the ideal GTA. If the observer's importance ratings matched the O*Net's perfectly, the difference-from-ideal-candidate score would be 0. Thus, low

scores indicate a concept of the ideal GTA very similar to the O*Net and higher scores indicate a concept of the ideal GTA different from the O*Net[6].

Procedure

All participants attended two sessions – an introductory session and a viewing session. In the introductory session, the general purpose of the study was explained, consent forms were signed, individual difference variables were measured, and the job description and rating scale were reviewed. In the viewing session, the job description and rating scale were reviewed again, the GTA interviews were viewed and rating forms filled out, and participants were debriefed. Nine pairs of introductory and viewing sessions were held, with attendance ranging from 15 to 52 participants per session.

Introductory Session. The purpose of the introductory session was to explain the purpose of the study, administer the general intelligence and ideal candidate measures, review the job analysis for a GTA, introduce the participants to the type of question relevant to their condition, and give participants an opportunity to use the rating scale.

I introduced the purpose of the study with the following statement:

> The purpose of this study is to see whether your accuracy in judging people being interviewed is related to the type of question asked. In addition, we are looking into whether your reasoning ability and your concept of the ideal candidate are related to accuracy. If you agree to participate, you must complete both of the sessions that you signed up for. Each session will last about 75 minutes.

Participants were given a consent form which described the study in greater detail and given an opportunity to ask any questions before signing and conveying the consent forms to research assistants (RAs). The WPT was administered under the standardized instructions included in the WPT Manual. The instructions shown at the beginning of the

ideal candidate survey were read to the participants (see Appendix C), who then completed the ideal candidate survey.

After the individual difference variables were measured, the written job description for a graduate teaching assistant was handed out and projected on a overhead screen (See Appendix A), which I introduced with the following statement:

> We are now handing out an abbreviated job description for a Graduate Teaching Assistant (which I'll refer to as GTA). Although abbreviated, this job description includes some of the typical tasks performed by GTAs and some of knowledge, skills, and abilities needed by GTAs.

I read several tasks and each of the knowledge, skills, and abilities to be assessed through the interview to the participants.

After reviewing the job description, written examples of two AI, PBDI, or SI questions related to KSAs of a GTA (as appropriate for each session), the rating scale related to each question, and three possible answers to each question were handed out and projected on an overhead screen and reviewed orally. None of these introductory session questions were asked during the video-recorded interview. The written instructions included on the rating scale form were read to participants. Two written examples of questions and three possible answers to each question were read to participants, who were given an opportunity to privately evaluate the written answers. The expected evaluation of a trained SME was then disclosed, as well as some of the key words or phrases to which an expert would have attended, which allowed participants to privately compare their evaluation to that of an expert. This training session lasted 15 to 20 minutes.

So that the training session would not unduly influence the participants' prototypes of the ideal teaching assistant, participants were asked to return to a second session in approximately two weeks in order to watch the interview videos. Prior to

departing, participants were told that the video recordings were portions of a longer interview in which additional information was gathered and that they would be asked to evaluate a limited number of qualities of the interviewees.

Viewing Session. Approximately two weeks after the introductory session, participants attended a second session in which the video recordings appropriate for their interview condition were viewed. In order to motivate the participants to be as accurate as possible, shortly after their arrival, participants were told:

> We are very hopeful that this study – and your participation in this study - will allow organizations to improve their interview techniques. So please do your best to listen carefully to the interviews, to take notes, and to assess the candidate's answers as accurately as you can.

In addition, two or three RAs positioned themselves in strategic locations throughout the room and monitored observers during the observation of the videos and, more importantly, while participants filled out their rating forms. In order to further motivate the participants to be as accurate as possible, the details of a lottery for the most accurate observers were disclosed at the beginning of the session. Among the participants in each condition, a drawing would be held among the top five most accurate raters and three would be eligible to receive a cash prize of $35.

The job description for a graduate teaching assistant was handed out and projected on a screen (see Appendix A). The general description for a GTA, several tasks, and the KSAs to be assessed through the questions asked in the video-recorded interviews were read aloud. As this was the first occasion on which participants would use the rating scale for the six questions to be asked and answered in the interview, the rating form was projected on a screen. The instructions were read to participants. Each of

the questions to be asked in the relevant condition and the rating scales were read aloud. The training lasted 15-20 minutes.

The three interviews related to each interview condition were viewed. In order to approximate real interview conditions, participants were encouraged to take notes, but were asked not to fill out the rating scale until the conclusion of the interview. The video recording was paused at the end of each interview and participants were given as much time as necessary to fill out the rating form. At the conclusion of the session, all rating evaluation forms were gathered, participants were debriefed on the purpose of the study, and told how they could find out about the results of the lottery.

Results

The means, standard deviations, and correlations among the study's variables are shown in Table 4. Accuracy was represented as the sum of the differences between a participant's rating of the answer to a question (scaled from 1 to 7) and the SME rating of the answer to a question. As such, lower accuracy scores represent greater accuracy. Across all conditions, general intelligence was significantly and positively associated with greater accuracy.

Difference-from-ideal-candidate represents the sum of the differences between a participant's importance value (scaled from 1 to 100) associated with a KSA and the importance value associated with the KSA on the O*Net. Higher scores indicate greater dissimilarity from the O*Net values. Across all conditions, accuracy was significantly and negatively correlated with greater dissimilarity between the participants' concept of the ideal GTA and the ideal GTA as portrayed on the O*Net.

Table 4.

Means, Standard Deviations, and Correlation Matrix

	M	SD	Accuracy	WPT
Accuracy	25.18	6.27		
WPT	24.16	5.44	-.16**	
Difference-from-Ideal-Candidate	686.13	159.41	.12*	-.10

Note. WPT = Wonderlic Personnel Test. Lower accuracy scores indicate greater accuracy. *p < .05. **p < .01

Hypothesis 1

The first hypothesis proposed that there would be a main effect for interview type, such that observer accuracy would be the least in the AI format and greatest in the SI format, with the degree of accuracy for the PBDI format falling between those two extremes. Table 5 shows the means and standard deviations for accuracy in each of the interview conditions. An ANOVA analysis indicated that there was a significant effect for interview type ($F(2, 298) = 17.15$, $p < .001$). Observer accuracy was greatest in the SI format, least in the PBDI format, with the AI format falling between those two extremes. A Bonferroni post hoc analysis, shown in Table 6, indicated that there were significant differences in accuracy between the PBDI and both the SI and AI conditions, but not between the SI and AI conditions. Thus, the first hypothesis was partially supported.

Table 5.

Means and Standard Deviations for Accuracy by Interview Condition.

Interview Condition	N	Mean Accuracy Score	Std. Deviation of Accuracy Score
SI	107	23.38	5.75
AI	99	24.34	5.43
PBDI	95	28.08	6.67

Note. Lower accuracy scores indicate greater accuracy.

Table 6.

Post Hoc Comparisons Between Interview Conditions.

(I) Condition	(J) Condition	Mean Difference (I-J)	Effect Size
AI	SI	.96	.17
	PBDI	-3.74*	-.62*
SI	PBDI	-4.71*	-.76*

*$p < .05$.

Hypothesis 2

The second hypothesis proposed that general intelligence would moderate the relationship between accuracy and interview type, such that general intelligence would be positively related to accuracy under the AI format, would not be related to greater accuracy under the SI format, with the relationship between intelligence and accuracy under the PBDI format falling in between these two extremes. The correlations between accuracy and intelligence by interview condition are shown in Table 7. There was a significant positive relationship between accuracy and intelligence (a negative relationship between inaccuracy and intelligence) in the SI condition.

Table 7.

Correlations between Accuracy and Intelligence by Interview Condition

Condition	*r*
AI	-.04
SI	-.30**
PBDI	-.17

**p < .01.

As Table 8 shows, the first step of the regression analysis also indicated a significant positive, but small, relationship between intelligence and accuracy (i.e., a significant negative relationship between general intelligence and inaccuracy) and a significant effect for the PBDI condition. However, the second step of the moderated regression analysis did not indicate a significant interaction effect between general intelligence and interview type (R^2 = .13, (F (3, 297) = 15.0, p < .001) for Step 1. Change R^2 = .01, (F (2, 295) = 1.505) for Step 2). Thus, Hypothesis 2 was not supported.

Table 8.

Hierarchical Regression Results for General Intelligence, Interview Condition, and Accuracy

Variables	B	Std. Error	β
Step 1			
Constant	29.12	1.64	
WPT	-.20	.06	-.17**
SI	-1.13	.82	-.09
PBDI	3.72	.84	.28**
Step 2			
Constant	25.49	2.91	
WPT	-.05	.12	-.04
SI	5.44	3.88	.42
PBDI	7.23	3.90	.54
SI * WPT	-.27	.16	-.51
PBDI * WPT	-.15	.16	-.27

Note. The AI condition served as the control variable for purposes of dummy coding.
**$p < .01$.

Hypothesis 3

The third hypothesis proposed that the interviewer's prototype of the ideal candidate would moderate the relationship between interviewer accuracy and interview structure, such that the relationship between the interviewer's difference-from-ideal-candidate and accuracy would be negative and greatest in the AI condition, and would not be related to accuracy in the SI condition, with the PBDI format falling in between these two extremes. The correlations between accuracy and difference-from-ideal-candidate are shown in Table 9. None of the relationships was significant in any of the interview conditions.

Table 9.

Correlations between Accuracy and Difference-from-Ideal-Candidate by Interview Condition.

Condition	*r*
AI	-.01
SI	.17
PBDI	.19

As Table 10 shows, in the first step of the regression analysis, there was significant, but small, negative relationship between difference-from-ideal-candidate and accuracy (or a significant small positive relationship between difference-from-ideal-candidate and inaccuracy), with a significant main effect for the PDBI condition. The second step of the moderated regression analysis did not indicate a significant interaction effect between difference-from-ideal-candidate and interview type ($R^2 = .12$, ($F (3, 297) = 13.02$, $p < .001$) for Step 1. Change $R^2 = .01$, ($F (2, 295) = 1.35$) for Step 2). Thus, Hypothesis 3 was not supported.

Table 10.

Hierarchical Regression Results for Difference-from-Ideal-Candidate, Interview Condition, and Accuracy.

Variable	B	Std. Error	β
Step 1			
Constant	21.17	1.63	
Difference-from-Ideal-Candidate	.00	.00	.12*
SI	-.80	.83	-.06
PBDI	3.80	.85	.28**
Step 2			
Constant	24.56	2.64	
Difference-from-Ideal-Candidate	-.00	.00	-.01
SI	-5.51	3.69	-.42
PBDI	-1.75	3.75	-.13
SI * Difference-from-Ideal-Candidate	.00	.01	.42
PBDI * Difference-from-Ideal-Candidate	.00	.01	.36

Note. The AI condition served as the control variable for purposes of dummy coding. *p < .05. **p < .01.

Discussion

Summary of Findings

Most comparative studies on the interview have examined differences in criterion validity and relationships with key variables across interview conditions (Conway & Peneno, 1999; Huffcut et al., 2001, Pulakos & Schmitt, 1995, Latham & Skarlicki, 1995). Van Iddekinge et al. (2004) examined differences in interviewer validity using a structured past-behavioral interview and a behaviorally anchored dimension-based rating scale with respect to performance evaluations. Few studies have examined differences in interviewer accuracy across interview conditions, with the exception of Mauer (2002) and Jako (1991). Mauer compared a behavioral SI scale to a conventional Likert scale and found the SI scale to result in greater accuracy and inter-rater agreement. Jako compared interviewer accuracy between a structured interview (comprised of attitudinal, situational, and past-behavioral questions) and an unstructured interview in assessing overall applicant-organization fit regarding personality traits and did not find significant differences in interviewer accuracy across conditions. As far as I know, this is the only study to examine observer accuracy across types of interview questions, while holding a question-based rating scale constant across conditions. Given that this is an initial investigation of this issue, perhaps it is not surprising that the study's results are not readily interpretable.

The first hypothesis proposed that there would be a significant positive relationship between interview structure and accuracy, and this was partially supported. The highly structured SI, in which the interviewees respond to a situational dilemma, without follow-up questions, resulted in the highest degree of observer accuracy. Unexpectedly, the next most accurate condition was the AI condition, although observer accuracy did not differ significantly from observer accuracy in the SI condition. The PDBI resulted in the least observer accuracy and differed significantly from both the SI and AI.

One of the ways in which both the SI and AI differed from the PBDI was that follow-up probes were not utilized in the SI and AI. Follow-up probes introduce a conversational aspect to the interview and may make it more difficult for observers to abstract the KSA-related information. Perhaps it was this conversational element – also an essential feature of the typical unstructured interview – which influenced the observers' accuracy.

Other than the presence of follow-up questioning, given that most other factors were held constant, one may look to the content of the questions and answers for other possible effects on accuracy. The most obvious difference is that the PBDI asks the interviewee to recount a prior experience and, based on my own observation, the answers tend to take on a story-like, this-happened-to-me quality, in which nuanced details are included. This story-like quality may make the KSA-relevant content less accessible than the content of AI answers, in which the interviewee explains his or her preferences or opinions, and the content of SI answers, in which the presentation of a particular dilemma, by its very nature, narrows the possible answers to some degree.

The second hypothesis proposed an interaction effect between intelligence and structure, which was not supported. It appears that more intelligent observers were only slightly more accurate across conditions, although there was a significant moderate relationship between intelligence and accuracy in the SI condition. And although the intelligence-by-condition interaction effect was not found, and we cannot conclude that there were significant differences in this regard across conditions, I would not have expected intelligence to be positively related to accuracy in the most structured, SI condition.

The third hypothesis proposed a moderating effect for the dissimilarity of the observer's concept of the ideal candidate from that of the O*Net on the relationship between structure and accuracy. This hypothesis was also not supported. Although there was a significant, but small, main effect for difference-from-ideal-candidate, after controlling for interview condition, there was no support for an interaction effect. Contrary to prior studies regarding the biasing influence of prototypes on accuracy in performance ratings (Nathan & Lord, 1983; Padgett & Ilgen, 1989), this study found only a weak relationship between difference-from-ideal-candidate and accuracy.

Given that the main effects of intelligence and difference-from-ideal candidate were small (Cohen, 1988), it is possible that this study had insufficient power to detect the interaction effect. Based on the sample size and effect size, the power of this study to detect the interaction was approximately .40, while .50 is considered to be a minimal power level and .80 is considered adequate.

Setting aside the question of sufficient power, other features of the interviews may have reduced the possibility of finding an interaction effect between intelligence and

interview structure and the observer's concept of the ideal candidate and interview structure. In particular, using a typical SI rating scale and holding it constant across interview conditions, as well as the extent of the training, may have compensated for a relationship between intelligence, structure, and accuracy or difference-from-ideal-candidate, structure, and accuracy. Mauer (2002) found evidence that a behaviorally-anchored rating scale resulted in a higher degree of rater accuracy than a structured conventional scale, with little difference in accuracy between experts and non-experts. By keeping the rating scale constant across interview conditions and asking observers to rate each answer, in effect, this study added an additional element of structure to the traditional AI and PBDI formats, which may have compensated for the effects of intelligence and difference-from-ideal-candidate which may otherwise have been found.

Ever since Campion et al.'s (1997) identification of 15 components of structure, as well as Huffcut & Arthur's meta-analytical finding that, beyond a certain point, additional structuring may not increase criterion validity, the question naturally arises as to which elements of structure are most important. In addition, researchers have begun to ask which individual difference variables contribute to better interviewer judgments (Dibpoye, 1992; Graves, 1993; Van Iddekinge, 2004). Although inconclusive, this study suggests that a question-based behaviorally-anchored rating scale may provide sufficient structure such that the role of certain individual difference variables may be reduced.

Limitations

Due to the experimental setting, this study examined a limited number of features of typical interviews conducted in an organizational setting. In a typical unstructured

interview, for example, not only is the interviewer asking questions and listening to answers, but he or she is also trying to think of the next question to ask. Given that the AI questions were standardized and no probes were utilized, the AI condition was as highly structured as the SI condition, which may also have contributed to the results. And although the PBDI in this study included follow-up probes, the degree of probing in a real interview could be more extensive, depending on the richness of the candidate's answer. In addition, the interviews shown in this study were much shorter than the typical interview conducted in an organizational setting. In the process of altering question responses so that each candidate reflected a targeted level of a KSA, it is possible that the responses also became more transparent. In short, the greater experiential complexity of a less structured and longer interview could easily result in a greater role for the interviewer's intelligence and concept of the ideal candidate.

Although this study attempted to interject a certain amount of gravity to the rating situation though the presence of RAs, and although the participants seemed motivated by the possibility of winning the lottery, the interviewer's motivations in an organizational setting undoubtedly would reflect multiple and more significant motivations. For example, at the conclusion of the interview process, a choice would be made regarding to whom to offer the position which could have repercussions for the interviewer (How would this person be to work with?), the organization (Would this person produce?), and the interviewee (Is the ultimate choice fair?). The more complex motivational setting could also result in individual difference variables, such as interview intelligence and concept of the ideal candidate, influencing the decision process.

Also related to the experimental setting is the nature of the participant pool. Perhaps the results would have been different with a group of older participants, with greater work or interviewing experience. However, Eder and Buckley (1988) have also argued that student participants can yield meaningful information regarding "process issues (i.e., observing and rating behavior), but not in research concerned with content issues (e.g., making selection decisions)" (p. 98). As previously noted, the sample size in this study may have been a limiting factor.

Directions for Future Research

This study kept the question-based behaviorally-anchored rating scale constant across interview conditions and manipulated the interview content in terms of the nature of the questions and the answers. If it was the additional structure provided by the consistent use of the rating scale that resulted in a minimal role for intelligence and concept of the ideal candidate with respect to accuracy, an interesting follow-up experiment would be to hold the interview content constant and to change the rating scale. Based on Huffcut and Arthur's (1994) taxonomy of structure, the rating scales could range from an overall rating, to dimensional ratings, to the question-by-question rating used in this study. Each of those three rating scales also could be tested with and without behavioral anchors.

In addition to manipulating question and answer content, this study also manipulated the use of follow-up probes. Given that the only interview condition for which accuracy was significantly different was the PBDI, which selectively utilized follow-up probes, another possibility for future research is to manipulate the extent of the

questions and answers between the interviewer and interviewee. For example, one could examine the effects on accuracy of an interview condition in which a candidate gives a lengthy but complete answer to interview conditions in which the interviewer has to work harder to elicit the same information by asking additional questions. Perhaps a more challenging aspect of content, suggested by this study, is that accuracy also may be affected by the way in which information is delivered. Is it the case, for example, that answers to past-behavioral questions tend to contain more embellishments than answers to attitudinal or situational questions – in effect, contain more distracters – which interfere with an interviewer's ability to assess the KSA-related content?

Conclusion

This study proposed and examined a main effect for interview type on accuracy, an interaction effect between general intelligence and structure on accuracy, and an interaction effect between an observer's concept of the ideal candidate (as compared to that portrayed within the O*Net) and structure on accuracy. Although a significant main effect was found for interview type, no interaction effects were observed. Due to the degree of control exercised with respect to common interview features, the null results for the interaction hypotheses are suggestive of effects for the type of rating scale utilized, the use of follow-up probes, or the nature of the answers elicited by some types of questions.

Although no firm conclusions can be drawn, the results of this study suggest that the use of a behaviorally-anchored rating scale, derived from KSA-related critical incidents, as well as rating each individual question may offset, to some extent, the

effects of individual differences variables related to rater accuracy, such as intelligence and concept of the ideal candidate. In addition, although interviewers and interviewees may resist the use of highly structured interviews, which lack the spontaneity and exchange provided by a conversation, the results of this study suggest that the inclusion of conversation, as well as answer content, may make it more difficult for observers to discern KSA-related information through the interview.

References

Arvey, R.D., & Campion, J.E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology, 35*, 281-322.

Bernardin, H.J., & Russell, J.E. (1998). *Human Resource Management: An Experimental Approach (2nd ed.)*. Boston: Irwin/McGraw Hill.

Borman, W.C. (1978). Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology, 63,* 135-144.

Borman, W.C. (1979a). Individual differences correlates of accuracy in evaluating others' performance effectiveness. *Applied Psychological Measurement, 3,* 103-115.

Borman, W.C. (1979b). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64,* 410-421.

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology, 50,* 655-702.

Carlson, R.E. & Mayfield, E.C. (1967). Evaluating interview and employment application data. *Personnel Psychology, 20,* 441-460.

Chapman, D.S., & Zweig, D.I. (2005). Developing a nomological network for interview structure: Antecedents and consequences of the structured selection interview. *Personal Psychology, 58*, 673-702.

Cohen, J. (1988). *Statistical Power Analysis for Behavior Science* (2nd ed.). Hillsdale, NJ: Erlbaum.

Conway, J.M. & Peneno, G.M. (1999). Comparing structured interview question types: Construct validity and applicant reactions. *Journal of Business and Psychology, 13,* 485 – 506

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.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* New York: Holt, Rinehart, and Winston.

Cronbach, L.J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin, 52,* 177-203.

Dalessio, A., & Imada, A.S. (1984). Relationships between interview selection decisions and perceptions of applicant similarity to an ideal employee and self: A field study. *Human Relations, 37,* 67-80.

James, L.R., Demaree, R.G., & Wolf, Gerrit. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69,* 85-98.

Dipboye, R.L. (1992). *Selection Interviews: Process Perspectives.* Cincinnati, OH: South-Western Publishing Co.

Dipboye, R.L. (1997). Structured selection interviews: Why do they work? Why are they underutilized? In N. Anderson & P Herriot (Eds.), *International Handbook of Selection and Assessment (455-473).* New York: John Wiley & Sons.

Dipboye, R.L., Gaugler, B.B., Hayes, T.L., & Parker, D. (2001). The validity of unstructured panel interviews: More than meets the eye? *Journal of Business and Psychology, 16,* 2001

Dougherty, T.W., Ebert, R.J., & Callender, J.C. (1986). Policy capturing in the employment interview. *Journal of Applied Psychology, 71*, 9-15.

Dreher, G.F., Ash, R.A., & Hancock, P. (1988). The role of the traditional research design in underestimating the validity of the employment interview. *Personnel Psychology, 41,* 315-327.

Eder, R.W. & Buckley, M.R. (1988). The employment interview: An interactionist perspective. *Research in Personnel and Human Resources Management, 6,* 75-107.

Fiske, S.T., & Taylor, S.E. (1991). *Social Cognition (2nd ed.).* New York: McGraw Hill.

Flanagan, J.C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358.

Gaugler, B.B., & Thornton, G.C. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74,* 6121-618.

Ghiselli, E.E. (1966). The validity of the personnel interview. *Personnel Psychology, 19*, 389-394.

Graves, L.M. (1993). Sources of individual differences in interviewer effectiveness: A model and implications for future research. *Journal of Organizational Behavior, 14,* 349-370.

Graves, L.M. & Karen, R.J. (1992). Interviewer decision processes and effectiveness: An experimental policy-capturing investigation. *Personnel Psychology, 45*, 313-340.

Graves, L.M. & Karren, R.J. (1999). Are some interviewers better than others? In R.W. Eder & M.M. Harris (Eds.). *The Employment Interview Handbook.* Thousand Oaks, CA: Sage Publications.

Guion, R.M. (1987). Changing views for personnel selection. *Personnel Psychology, 40,* 199-213.

Harris, M.M. (1989). Reconsidering the employment interview: A review of recently literature and suggestions for future research. *Personnel Psychology, 42,* 691-726.

Heneman, H.G. III, Schwab, D.P., Huett, D.L., & Ford, J.J. (1973). Interviewer validity as a function of interview structure, biographical data, and interviewee order. *Journal of Applied Psychology, 60,* 748-753.

Hovland, C.I., & Wonderlic, E.F. (1939). Prediction of industrial success from a standardized interview. *Journal of Applied Psychology, 23,* 537-546.

Huffcutt, A.I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79,* 184-190.

Huffcutt, A.I., Weekley J., Wiesner, W.H., DeGroot, T., & Jones, C. (2001). Comparison of situational and behavior description interview questions for higher-level positions. *Personnel Psychology, 54,* 619-644.

Hunter, J.E. & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72-98.

Jako, RA. (1991). An investigation of the predictor-criterion relationships comprising employment interview validity. Unpublished doctoral dissertation, Colorado State University, Fort Collins.

Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interview. *Journal of Applied Psychology, 67,* 577-580.

Janz, T. (1989) Patterned behavior description interview. In R.W. Eder & G.R. Ferris (Eds.), *The employment interview: Theory, research, and practice (158-168).* Newbury Park, CA: Sage.

Janz, T., Hellervik, L., & Gilmore, D.C. (1986). *Behavior Description Interviewing: New, Accurate, Cost-effective.* Newton, MA: Allyn and Bacon, Inc.

Jeanneret, P.R., & Strong, M.H. (2003). Linking O*Net job analysis information to job requirement predictors: An O*Net application. *Personnel Psychology, 56,* 465-492.

Latham, G.P. (1989). The reliability, validity, and practicality of the situational interview. In R.W. Eder & G.R. Ferris (Eds.), *The employment interview: Theory, research, and practice (169-182).* Newbury Park, CA: Sage

Latham, G.P., Saari, L.M., Pursell, E.D., & Campion, M.A. (1980). The situational interview. *Journal of Applied Psychology, 65,* 422-427.

Latham, G.P. & Sue-Chan, C. (1999). A meta-analysis of the situational interview: An enumerative review of reasons for its validity. *Canadian Psychology, 40,* 56-67.

Lievens, F., & De Paepe, A. (2004). An empirical investigation of interviewer-related factors that discourage the use of high structure interviews. *Journal of Organizational Behavior, 25,* 29-46.

Locke, E.A. & Latham, G.P. (1984). *Goal Setting: A Motivational Technique that Works.* Englewood Cliffs, NJ: Prentice Hall.

Lord, R.G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology, 70,* 86-71.

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

Maurer, S.D., Sue-Chan, C., Latham, G.P. (1999). The situational interview. In R.W. Eder & M.M. Harris (Eds.), *The Employment Interview Handbook (159-177).* Newbury Park, CA: Sage.

Mayfield, E.C., & Carlson, R.E. (1966). Selection interview decisions: First results from a long-term research project. *Personnel Psychology, 19,* 41-53.

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.

McMurry, R.N. (1947). Validating the patterned interview. *Personnel, 23,* 263-272

Motowidlo, S.J. (1999). Asking about past behavior versus hypothetical behavior. In R.W. Eder & M. M. Harris (Eds.), *The Employment Interview Handbook (179-190).* Newbury Park, CA: Sage.

Motowidlo, S.J. (1986). Information processing in personnel decisions. *Research in Personnel and Human Resources Management, 4,* 1-44.

Murphy, K.R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W.K. (1982). Relationship between observational accuracy and accuracy in evaluating performance. *Journal of Applied Psychology, 67,* 320-325.

Murphy, K.R., & Myors, B. (1998). *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Testing.* Mahwah, NJ: Lawrence Erlbaum Associates.

Nathan, B.R., & Lord, R.G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. *Journal of Applied Psychology, 68,* 102-114.

O*Net Summary and Details Report for: 25-1191.00 – Graduate Teaching Assistants. Retrieved April 27, 2005, from http://online.onetcenter.org/link/summary/25-1191.00.

Padgett, M.Y., & Ilgen, D.R. (1989). The impact of ratee performance characteristics on rater cognitive processes and alternative measures of rater accuracy. *Organizational Behavior and Human Decision Processes, 44,* 232-260.

Pulakos, E.D. & Schmitt N. (1995). Experience-based and situational interview questions: Studies of validity. *Personnel Psychology, 48,* 289-308.

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 validity of a structured interview. *Personnel Psychology, 49,* 85-102.

Rowe, P.M. (1963). Individual differences in selection decisions. *Journal of Applied Psychology, 47,* 304-307.

Rowe, P.M. (1984). Decision processes in personnel selection. *Canadian Journal of Behavioral Science, 16,* 326-337

Schmidt, F.L., & Zimmerman, R.D. (2004). A counterintuitive hypothesis about employment interview validity and some supporting evidence. *Journal of Applied Psychology, 89,* 553-561.

Smither, J.W., & Reilly, R.R. (1987). True intercorrelations among job components, time delay in rating, and rater intelligence as determinants of accuracy in performance ratings. *Organizational Behavior and Human Decision Processes, 40,* 369-391.

Taft, R. (1955). The ability to judge people. *Psychological Bulletin, 52,* 1-23.

Terpstra, D.E., & Rozell, E.J. (1997). Why some potentially effective staffing practices are seldom used. *Public Personnel Management, 26,* 483-495.

Thornton, G.C. III (1992). *Assessment Centers in Human Resource Management.* Reading, MA: Addison-Wesley Publishing Company.

Taylor, P. & Small, B. (2000). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and behavior description interview questions. *Journal of Occupational and Organizational Psychology, 75,* 277-294

Valenzi, E., & Andrews, I.R. (1973). Individual differences in the decision process of employment interviewers. *Journal of Applied Psychology, 58,* 49-53.

Van der Zee, K.I., Bakker, A.B., & Bakker, P. (2002). Why are structured interviews so rarely used in personnel selection? *Journal of Applied Psychology, 87,* 176-184.

Van Iddekinge, C.H., Sager, C.E., Burfield, J.L. & Heffner, T.S. (2004). A closer look at differences in interviewer validity and reliability. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychologists, Chicago, IL.

Vernon, P.E. (1933). Some characteristics of the good judge of personality. *Journal of Social Psychology, 4,* 42-57.

Wagner, R. (1949) The employment interview: A critical summary. *Personnel Psychology, 2,* 17-46.

Wiesner, W.H., & Cronshaw, S.F. (1988). A meta-analytical 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.

Wilk, S., & Cappelli, P. (2003). Understanding the determinants of employer use of selection methods. *Personnel Psychology, 56,* 103-123.

Wonderlic Personnel Test & Scholastic Level Exam User's Manual. Libertyville, IL: Wonderlic, Inc.

Wright, P.M., Lichtenfels, P.A., & Pursell, E.D. (1989). The structured interview: Additional studies and meta-analyses. *Journal of Occupational Psychology, 62,* 191-199

Zedeck, S., Tziner, A., & Middlestadt, S.E. (1983). Interviewer validity and reliability: An individual analysis approach. *Personnel Psychology, 36*, 355-370.

Appendix A

Job Description - Graduate Teaching Assistant

Assist department chairperson, faculty members, or other professional staff members by performing teaching or teaching-related duties, such as teaching lower level courses, developing teaching materials, preparing and giving examinations, and grading examinations or papers. Graduate assistants must be enrolled in a graduate school program.

Tasks

- Evaluate and grade examinations, assignments, and papers, and record grades.
- Lead discussion sections, tutorials, and laboratory sections.
- Teach undergraduate level courses, if necessary
- Develop teaching materials such as syllabi, visual aids, answer keys, supplementary notes, and course websites.
- Attend lectures given by the instructor whom they are assisting.
- Complete laboratory projects prior to assigning them to students so that any needed modifications can be made.
- Copy and distribute classroom materials.
- Demonstrate use of laboratory equipment, and enforce laboratory rules.
- Inform students of the procedures for completing and submitting class work such as lab reports.
- Meet with supervisors to discuss students' grades, and to complete required grade-related paperwork.

Knowledge

Education and Training:

- Demonstrates knowledge of principles and methods for designing course material.
- Demonstrates knowledge of principles and methods for assessing learning.

English Language: Knowledge of the English language, including spelling, composition, and grammar.

Skills

Instructing: Teaching others how to do something.

Critical Thinking: Use of logic and reasoning to compare and contrast alternative solutions or approaches to problems.

Abilities

Oral Expression: The ability to communicate information and ideas through speaking

Speech Clarity: The ability to speak clearly (good enunciation, for example).

Appendix B

Interview Questions and Rating Scales

Attitudinal Question: After presenting lecture material to students regarding a challenging assignment, how do you like to assess their grasp of the material?

Situational Question: Suppose that you just presented lecture material related to solving a challenging assignment. Some of your students seem to prefer for you to present several examples of this type of assignment in class, while others seem to prefer to independently work on the assignment outside of the classroom. How would you assess whether the students have understood the lecture material?

Past-behavioral question: Would you describe for me a recent lecture in which you presented a challenging assignment to your students?

Possible Follow-Up Probe:
How did you assess whether the students understood?

Education and Training						
• Demonstrates knowledge of principles and methods for assessing learning.						
Low		Moderate			High	
1	2	3	4	5	6	7
After presenting lecture material, gives assignment related to lecture material		After presenting lecture material, provides class time to discuss the assignment			After presenting lecture material, provides class time to work on the assignment and receive feedback from the instructor or peers	

Interview Questions and Rating Scale

Attitudinal Question: What sorts of portfolio assignments do you like to design?

Situational Question: Suppose your class is comprised of students ranging from the very bright to those less gifted. You sometimes feel that the brightest are bored with the course material, but that other students are struggling to keep up. What sort of portfolio assignment would you design for this class?

Past-behavioral question: Tell me about the best portfolio assignment you ever designed.

Possible Follow-Up Probe:
What was the students' reaction?

Education and Training		
• Demonstrates knowledge of principles and methods for assessing learning.		
Low	Moderate	High
1 2	3 4 5	6 7
Portfolio assignments require the grasp of the most challenging prior work and homework material	Portfolio assignments require the grasp of material similar to prior work and homework assignments, but also require the student to go beyond the prior work and homework material	Portfolio assignments require the grasp of material very similar to prior work and the homework material

Interview Questions and Rating Scale

Attitudinal Question: When you are teaching students how to do something, for example, write an essay, do you think it's important to make the subject interesting for them?

Situational Question: Suppose you are teaching your students how to do something, for example, write an essay. Although some members of the class are asking questions and seem quite interested, several members of the class seem bored. What would you do?

Past-behavioral question: Tell me about the last time you were instructing your students on how to do something - for example, write an essay - and some of the students seemed bored.

Possible Follow-Up Probes:
So what type of [examples, games, stories] did you explore?
What was the students' reaction?

Skills						
• Teaching others how to do something.						
Low		Moderate			High	
1	2	3	4	5	6	7
When teaching others how to do something, does not show a real-world application		When teaching others how to do something, tells students that what they are learning relates to real-world problems, but does not show them how			When teaching others how to do something, shows how the topic could be used in a real-world application	

Interview Questions and Rating Scale

Attitudinal Question: How do you generally respond to students seeking assistance with assignments during your office hours?

Situational Question: Suppose you had just delivered a series of lectures on how to do something, for example, write a particular type of essay. You worked very hard on the lectures and presented several examples very similar to the assignment. You feel that the material could not have been more clear. On a particularly busy day for you, a student comes to you during your office hours and says: "I've gone over all my notes and I still can't seem to do the assignment." What would you do?

Past-behavioral question: Tell me about the most difficult time you have had with a student seeking help with an assignment during your office hours.

Possible Follow-Up Probe:
What was his or her reaction?

Skills		
• Teaching others how to do something.		
Low	Moderate	High
1 2	3 4 5	6 7
Asks students seeking assistance with assignment during office hours to review the class material again Perhaps suggests alternative sources of help	Assists students seeking assistance with the assignment during office hours by going over examples of how the instructor or others have done the type of assignment	Assists students seeking assistance with the assignment during office hours by working through parts of the assignment with them

Interview Questions and Rating Scale

Attitudinal Question: If more than one method exists for approaching an assignment, what is your approach in your teaching?

Situational Question: Suppose that you are presenting several ways to approach an assignment to your class. Some writers in your field believe that there is one best way to approach this type of assignment. Other writers in your field believe that the way to approach the assignment should be determined by the circumstances. How would you proceed?

Past-behavioral question: Tell me about your most challenging instance of presenting different approaches to an assignment.

Possible Follow-Up Probes:
So is there one method that you endorse?
Do you allow the students to decide on the approach?
What was the students' reaction?

Skills		
• Use of logic and reasoning in order to compare and contrast alternative solutions or approaches to problems.		
Low	Moderate	High
1 2	3 4 5	6 7
Presents alternative approaches to assignment, but strongly endorses one method, without setting clear standards	Presents alternative approaches to the assignment, and evaluates some as better than others, but without setting clear standards	Presents alternative approaches to the assignment, and establishes standards by which to judge alternative solutions. Using the standards, compares and contrasts alternatives

Interview Questions and Rating Scale

Attitudinal Question: What is your reaction when one or two students perform an assignment very differently than was presented in class?

Situational Question: Suppose you have asked your students to present their approaches to an assignment. One or two students present an approach to the assignment that is correct, but very different than what was previously discussed in class. On the one hand, you are concerned that discussing the students' approach may confuse the class. On the other hand, you do not want to discourage these students. What would you do?

Past-behavioral question: Tell me about a time when one or two students performed an assignment very differently than was presented in class.

Possible Follow-Up Probes:
So what did you do with those students?
What was the students' reaction?

Skills		
• Use of logic and reasoning to compare and contrast alternative solutions or approaches to problems.		
Low	Moderate	High
1 2	3 4 5	6 7
When a student presents a very different way of approaching an assignment than was shown in class, regards the student's approach as too confusing to the class and steers the class away from that approach	When a student presents a very different way of approaching an assignment than was shown in class, notes that the student's approach is one way to do the problem, but focuses on one preferred approach	When a student presents a very different way of approaching an assignment than was shown in class, works through the student's approach with the class and compares and contrasts to alternative approaches

Rating Scales for KSAs Directly Observed during the Interview

Knowledge						
• Knowledge of the English language, including spelling, composition, and grammar.						
Low		Moderate			High	
1	2	3	4	5	6	7

Abilities						
The ability to communicate information and ideas through speaking.						
Low		Moderate			High	
1	2	3	4	5	6	7

Abilities						
• The ability to speak clearly (good enunciation, for example).						
Low		Moderate			High	
1	2	3	4	5	6	7

Appendix C

The Ideal Candidate

Below you will find a list of qualities commonly used to characterize graduate teaching assistants. Please use this list to tell me how important these qualities are to your *ideal* graduate teaching assistant (GTA). Your ideal GTA is your concept of the perfect GTA.

If a quality is very important or essential to your concept of the perfect GTA, you should give it the highest score of 100. If a quality is not important at all to your concept of the perfect GTA, you should give it the lowest score of 0. If a quality is moderately important – helpful, but not essential - you might want to give it a score close to 50.

Please write a number between 0 and 100 in the line to the left of the item. You should feel free to use numbers anywhere between 1 and 100 – 38 is a perfectly good number.
Remember:

100 means that the quality is extremely important
75 means that the quality is important
50 means that the quality is moderately important
25 means that the quality is somewhat important
0 means that the quality is not important at all

Graduate Teaching Assistant

Assist department chairperson, faculty members, or other professional staff members by performing teaching or teaching-related duties, such as teaching lower level courses, developing teaching materials, preparing and giving examinations, and grading examinations or papers. Graduate assistants must be enrolled in a graduate school program.

_____ 1) **Communication of Media:** Knowledge of media production.

_____ 2) **Education and Training:** Demonstrates knowledge of principles and methods for designing lectures.

_____ 3) **Education and Training:** Demonstrates knowledge of principles and methods for assessing learning.

_____ 4) **English Language:** Knowledge of the English language, including spelling, composition, and grammar.

_____ 5) **Production and Processing:** Knowledge of production processes and quality control.

_____ 6) **Instructing:** Teaching others how to do something.

_____ 7) **Judgment and Decision Making:** Consider the relative costs and benefits of actions.

_____ 8) **Coordination:** Adjusting actions in relation to others' actions.

_____ 9) **Critical Thinking:** Use of logic and reasoning in order to compare and contrast alternative solutions or approaches to problems.

_____ 10) **Social Perceptiveness:** Being aware of others' reactions.

_____ 11) **Negotiation:** Bringing others together and trying to reconcile differences.

_____ 12) **Management of Financial Resources:** Determining how money will be spent.

_____ 13) **Reading Comprehension:** Understanding written sentences and paragraphs.

_____ 14) **Troubleshooting:** Determining causes of operating errors and deciding what to do.

_____ 15) **Oral Expression:** The ability to communicate information and ideas through speaking.

_____ 16) **Memorization:** The ability to remember information.

_____ 17) **Far Vision:** The ability to see details at a distance.

_____ 18) **Speech Clarity:** The ability to speak clearly (good enunciation, for example).

_____ 19) **Time Management:** Managing one's own time and the time of others.

_____ 20) **Originality:** The ability to come up with unusual or clever ideas.

_____ 21) **Stamina:** The ability to exert yourself physically over long periods of time.

_____ 22) **Oral Comprehension:** The ability to listen to and understand information and ideas presented through spoken words.

_____ 23) **Written Expression:** The ability to communicate information and ideas in writing.

_____ 24) **Night Vision:** The ability to see under low light conditions.

_____ 25) **Reaction Time:** The ability to quickly respond (with hand, finger, or foot) to a signal (sound, light, picture).

If you think there are additional knowledge, skills, or abilities relevant to the job of Graduate Teaching Assistant, please write them in below and indicate the degree of importance you associated with the characteristic.

Additional Characteristics

_____ _____

_____ _____

_____ _____

_____ _____

_____ _____

_____ _____

_____ _____

_____ _____

Endnotes

[1] As various meta-analyses use different artifact corrections, upon the first occurrence of corrected validity coefficients from a particular meta-analysis, the specific artifact corrections will be noted. Subsequent references will not indicate the nature of the corrections.

[2] Very recently, Chapman and Zweig (2005) proposed a four-dimensional factor model of interview structure, which included: question consistency, evaluation standardization, question sophistication, and rapport building. It is yet to be determined whether interview researchers will embrace this model.

[3] Several court rulings in the 1970s have clarified that employers may use only job-related criteria in making personnel decisions. As such, in most cases, questions regarding age, marital status, and military service now would be considered illegal.

[4] After making this observation, Motowidlo argued against the claim that hypothetical behavior in future situations does in fact measure intentions. He argued that this is unlikely because: There is pressure on applicants to create a favorable impression in the interview, intentions may change over time, and respondents may have a lack of control over volitional intended behaviors.

[5] Huffcutt et al. used the acronym, BDI – perhaps indicating the lack of patterning.

[6] I investigated the possibility of reporting reliability indices for the accuracy and the difference-from-ideal-candidate difference scores. However, the standard calculation for a difference scores requires that there be some variance on both the rated scores and the true scores (Crocker & Algina, 1986). Given that I requested that the SMEs reach consensus in their identification of the level of the dimension reflected in an answer, and given that a single set of O*Net importance values was used in calculating the difference-from-ideal-candidate, the SME scores and the O*Net importance values displayed no variance. For this reason, reliability indices for these difference scores could not be calculated.