2011

Computer Adaptive Rating Scales (CARS) for the Employment Interview

Greg F. Schmidt
University of South Florida, business.account@gmail.com

Computer Adaptive Rating Scales (CARS)

for the Employment Interview

by

Greg F. Schmidt

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Psychology
College of Arts and Sciences
University of South Florida

Major Professor: Walter C. Borman, Ph.D.
Michael T. Brannick, Ph.D.
Stephen Stark, Ph.D.
Joseph A. Vandello, Ph.D.
Walter R. Nord, Ph.D.

Date of Approval:
June 17, 2011

Keywords: selection, interview, performance appraisal, rating format, adaptive rating

Table of Contents

List of Tables ii

Abstract iii

Chapter 1: Introduction 1
- History of Selection Interview Research 2
- Structure of a Selection Interview 4
- Using Interviews in Organizational Settings 5
- Interview Rating Scales 8
- Paired Comparisons 11
- Development and Evaluation of CARS 13
- Current Study 14
- Hypotheses 15

Chapter 2: Method 17
- Performance Domain 17
- Scale Development 18
  - Building the CARS Program 18
  - Development of Behavioral Statements 20
  - BARS Development 20
- Videotaped Interviews 21
- Job Performance Ratings 22
- Data Collection 22

Chapter 3: Results 25

Chapter 4: Discussion 30
- Theoretical Implications 32
- Applied Implications 36
- Limitations 37
- Summary 38

References 40

Appendices 47
- Appendix A: Examples of CARS Behavioral Statements 48

List of Tables

Table 1: Descriptives for Overall and Dimension-level ratings across conditions 25

Table 2: Correlations between BARS, CARS, and Supervisor Performance Ratings for Leadership Dimension 26

Table 3: Correlations between BARS, CARS, and Supervisor Performance Ratings for Teamwork Dimension 26

Table 4: Correlations between BARS, CARS, and Supervisor Performance Ratings for Planning & Organization Dimension 27

Table 5: Correlations between BARS, CARS, and Supervisor Performance Ratings for Drive Dimension 27

Table 6: Correlations between BARS, CARS, and Supervisor Performance Ratings for Total Performance 28

Table 6: Test for Equality of Dependant Correlations with one common variable 29

Abstract

This research investigates the effectiveness of computerized adaptive rating scales (CARS) in comparison to the relatively more common behavioral anchored ratings scales (BARS) format. The current study sought to extend the body of psychometric research of CARS while investigating its potential for use in the employment interview. Using 43 videotaped interviews and supervisor job performance ratings, and constructing a new task-performance based CARS, it was hypothesized that employment interview ratings produced using the CARS format would yield significantly higher predictive validity coefficients than those produced by the BARS format. Results showed that while interview ratings produced in the CARS format were predictive of supervisor job performance ratings, they were not significantly higher than ratings in the BARS format. Academic and applied implications are discussed.

Chapter 1: Introduction

Industrial and organizational psychologists have been studying the employment interview for more than 80 years in an attempt to improve our ability to identify those candidates who are most likely to succeed in a particular job (Dipboye & Gaugler, 1993; Latham, Saari, Pursell, & Campion, 1980; Posthuma, Morgeson, & Campion, 2002). Currently, the employment interview remains an important element of many organizations' selection process, often complementing other selection strategies such as prescreening, work samples, cognitive ability tests, and personality assessments- all used to narrow an applicant pool down to a select few qualified candidates.

Of all these options, the employment interview is probably the most commonly used (Harris, 1989; Posthuma, Morgeson, & Campion, 2002). In fact, one would be hard pressed to find an organization willing to hire an associate without some sort of pre-employment conversation. However, between scheduling, conducting, and evaluating each employment interview, many large organizations are forced to devote a great deal of their human and financial resources to adequately execute the process. To this end, applied I/O psychologists have spent much time and effort over the years ensuring this is time and money well spent, and that interviews are valid and reliable predictors of top organizational performers. In fact, over the past 80 years, there has been no shortage of research on employment interviews, focused on different elements such as interviewer training, interviewee characteristics, faking, interview format, among others (for

thorough reviews see: Arvey & Campion, 1982; Mayfield, 1964; Posthuma, Morgeson & Campion, 2002; Schmitt, 1976; Ulrich & Trumbo, 1965; Wagner, 1949).

The purpose of this study is to investigate how a particular structural element of the interview, the performance rating scale format, affects employment interview outcomes. Specifically, we will examine how employment interview ratings made using computer adaptive rating scales compare to ratings made using more traditional behaviorally anchored rating scale formats. The paper will begin with a summary of relevant research about the employment interview, and will then review research focused on the rating scales used to evaluate and compare interviewee's responses to questions. Based on this literature, we predict that hiring recommendations made using computer adaptive rating scales (CARS) format to evaluate verbal interview responses will be comparatively more valid than recommendations made using more traditional rating scales (e.g. behavioral anchored rating scales, or BARS).

History of Selection Interview Research

As mentioned previously, several major reviews of interview research summarize the immense wealth of available information on the employment interview, and provide an interesting glimpse into how experts in the field - and organizations – historically and currently view the interview (Arvey & Campion, 1982; Mayfield, 1964; Posthuma, Morgeson & Campion, 2002; Schmitt, 1976; Ulrich & Trumbo, 1965; Wagner, 1949). Many of the early reviews by Wagner (1949), Mayfield (1964), and Ulrich and Trumbo (1965) were rather pessimistic in their evaluations of the interview. I/O practitioners at the time tended to agree that the pre-hire interview used by many organizations was so plagued by problems that it had limited ability to accurately identify the right person for a

job. For instance, early research found that interviewers were unable to agree with each other on the best way to conduct the actual interview. Some interviewers felt their own unscripted, unstructured interview questions were most effective in gathering predictive information during the interview, while others argued for more structured and consistent interview questions. The research also showed that interviewers tended to disagree on how to evaluate the performance of the candidate during the actual interviewing process. Furthermore, early researchers concluded that irrelevant job performance information such as applicant appearance, attractiveness, and non-verbal cues (eye-contact, smiling, etc.) biased interviewer evaluations so much that it significantly undermined the predictive validity estimates of the interview (Pingatore, Dugoni, Tindale, & Spring, 1994; Dipobye, 1992). With so much negative press, many practitioners became wary of the interview as a pre-employment performance evaluation tool, citing a lack of return on the invested time and effort to interview, and an increased legal risk when used to hire new employees. Nonetheless, despite the worrisome climate surrounding the interview as a selection tool, managers understandably continued to insist on a face-to-face meeting with their future employees before extending an offer, perhaps due to the lack of alternative selection strategies.

Thankfully, more recent theoretical advances have shone a new positive light on the interview, and empirical advances have led to a more supportive and complex understanding of the selection interview. Beginning with Schmitt's (1976) review of the interview, a distinct change of opinion became noticeable in the field. In general, stronger empirical evidence demonstrated better psychometric properties of the employment interview and more favorable prediction of future job performance; thus, practitioners

and hiring managers alike became more confident in hiring decisions based on employment interviews. Subsequent review articles by Arvey and Campion (1982) and Harris (1989) argued that when certain structural procedures are followed, reliability and validity estimates of the employment interview were more favorable. Based on a thorough review of the available literature, the authors concluded that when organizations follow best practices in the development, implementation, and use of interviews to hire, the problems cited by earlier authors were minimized. In general, they found that when organizations based their interview questions on job analysis data, structured the interview format to keep interviewers systematic and objective across all applicants, and used multiple raters instead of just a single interviewer, more favorable interviewing outcomes were produced.

The promising empirical support for the selection interview has since continued, and the most recent reviews of the selection interview have been consistent with the notion that a structured interview can be a strong predictor of on-the-job performance (Judge, Higgins, & Cable, 2000; Posthuma, Morgeson, & Campion, 2002). These authors point out that the corrected validity coefficients from validity generalization studies of selection interviews are similar to those produced by cognitive ability tests and assessment centers- two selection strategies that have long been accepted for reliably and accurately selecting candidates into organizations (Schmidt & Hunter, 1998).

Structure of a Selection Interview

Today, there are many well documented guidelines for structuring an interview that help ensure its reliability and validity. At their most basic level, selection interviews are comprised of a conversation between the interviewer and interviewee, during which

the interviewer asks the interviewee a number of interview questions, and records his or her responses to allow for comparisons with other job candidates.

There are a few popular approaches for structuring an interview. The situational interview (sometimes called a future-oriented interview), first described by Latham et al. (1980), asks applicants to respond how they might behave in future hypothetical situations. An example situational interview question could be "What would you do if the work of a co-worker or subordinate was not up to expectations?" or "What would you do if the priorities on a project you were working on changed suddenly?" Motowidlo, Carter, Dunnette, Tippins, Werner, Burnett, and Vaughan (1992) introduced another type of structured interview, called the structured behavioral interview, in which all questions about past behavior are the same for all applicants. An example structured behavioral interview question could be "Describe a situation in which you were able to persuade someone to see things your way" or "Give me an example of a time when you set a challenging goal and were able to meet or achieve it". The McDaniel et al. (1994) meta-analysis found that situational or future-oriented interviews were somewhat more valid than behavioral interviews.

Using Interviews in Organizational Settings

Notwithstanding the advancements in employment interview theory, applied I/O psychologists continue to have difficulty convincing their applied clients to use structured interviews consistently and effectively. Despite the clear psychometric evidence that structured interviews are better predictors of job performance and do reduce rating errors, hiring managers still often resort to unstructured or informal conversations with candidates in which they "size up" the individual and make a subjective decision on their

"hire-ability". Recent popular press books such as Malcolm Gladwell's (2005) *Blink: the Power of Thinking Without Thinking* and Gerd Gigerenzer's (2007) *Gut Feelings: The Intelligence of the Unconscious* promote the idea that humans are effective intuitive decision makers. Without disputing these authors' claims of human ability to make snap judgments of others, there is strong evidence that intuitive selection decisions – the result of unstructured interviews – are unreliable and not valid predictors of future job performance (McDaniels et al., 1994; Schmidt & Hunter, 1998). In fact, the estimated interrater reliability of unstructured interview decisions is so low that judgments likely could never account for more than 10% of the observed variance in job performance (Conway, Jako, & Goodman, 1995).

One possible explanation for the stubborn insistence on intuition was the topic of a debate in the journal *Industrial and Organizational Psychology: Perspectives on science and practice.* In his focal article, Highhouse (2008) argues that the biggest failure of the field of I/O psychology is the inability to convince HR managers to use advanced selection tools. He goes on to theorize that this may be in part caused by the fact that managers believe it is possible to perfectly predict a candidate's job performance, and that selection tools (such as structured interviews) only impeded their ability to do so.

Researchers have consistently noted that significant variation exists between interviewers' abilities to accurately predict job performance, and have begun calling for the field to focus not just on the validity of the interview, but also the validity of the interviewer (Judge et al., 2000). Even when interviewers are given scientifically developed structured interview questions and rating scales, subjective evaluations can

lead to erroneous forecasts of job performance. As such, applied I/O psychologists often face a number of practical limitations in organizational interviewing situations that may undermine the adoption and effectiveness of structured employment interviews. Hiring managers, untrained in psychometrics, may believe that given the opportunity, they can choose the right person for the job without any chance of mistake. Consequently, I/O practitioners are consistently challenged to convince their clients that the structured interview is in fact better than an individual's intuition and subjective impression of a candidate.

One potential solution to this practical issue could be to construct the materials that aid interviewers during the interview process such that they minimize subjective interviewee performance appraisal, and 'force' an individual to produce more objective ratings to arrive at a hiring decision. Such a tool would limit a hiring managers' ability to influence the hiring decision with their own intuition. Of course, many of the aforementioned recommendations to improve interview outcomes (e.g. structuring the questions, conducting interviewer training, basing interview questions on job analysis data, etc.) are designed to do exactly that- minimize the influence of unreliable interviewer intuition in the employment decision. However, many of these approaches are not focused on the step in the interviewing process where a decision is actually made. Instead, many of these recommendations concern themselves with how interviews are developed (e.g. conducting job analyses, competency modeling) or the interview itself (e.g. interview guides). These strategies are certainly useful for improving the reliability and validity of interviews and should not be overlooked, but if we still allow opportunities for interviewers to make subjective decisions while they provide ratings,

our intentions to increase the reliability and validity of the interview may be limited. Some organizations have implemented interviewer skills training programs to instruct interviewers how to base their hiring decisions on objective interviewee data, but it is difficult to ensure an interviewer actually practices what they learn when interviewing.

To this end, we propose that another structural element of the interview- the rating scale format- may be used to encourage valid interview decisions, and encourage the interviewer to provide more objective ratings of interviewee performance before making the hiring decision. The remainder of this paper is focused on an effort to minimize the subjectivity in the employment decision by creating a selection decision aid to be used during the actual rating and decision-making process.

Interview Rating Scales

Over the years, the rating scale format is one structural component of the interview that has received relatively minimal attention concerning its impact on selection decisions. Rating scales are often provided to hiring managers to assist with organizing, recording, and evaluating interviewee responses to questions. Such interview rating scales are often standardized for all interviewers, and are designed to support accurate, valid, and reliable comparisons across job applicants. Over the years, there have been a few attempts by I/O psychologists to develop rating formats that were designed to reduce rating errors and improve the reliability and validity of interview outcomes (Landy & Farr, 1983; Murphy & Cleveland, 1995).

In 1963, Smith and Kendall introduced a new rating form known as behavioral anchored rating scales (BARS), a format that has become popular with I/O practitioners. BARS differ from simple likert-type scales by "anchoring" descriptive behavioral

statements at various places along the scale, providing the rater with specific behavioral examples that correspond to the numerical scale. Smith and Kendall believed that with descriptive examples, raters would be able to provide more objective and consistent evaluations of ratees.

Since then, a number of similar rating scales were introduced, many of them involving slight improvements to the BARS. For instance, Latham and Wexley (1981) developed behavior observation scales, and Blanz and Ghiselli introduced the mixed standard scale (1972). In 1979, Borman proposed behavior summary scales (BSS), followed by Kane (1986) who developed the performance distribution assessment method. Over the years, additional research focusing on the impact these different rating templates have on the reliability and validity of interview decisions has been conducted, ranging from the placement of the scale on the rating forms (Madden & Bourdon, 1964) to altering the number of response categories on the scale (Bendig, 1954). However, most of this research found that even when utilized correctly, the potential for rating scale format to reduce rating errors such as halo, leniency, severity, and central tendency, and to increase interview validity was minimal (e.g. Taylor & Wherry, 1951). Furthermore, whereas each format may have its benefits, research generally agrees that the rating format accounts for a relatively small amount of variance in psychometric quality of the rating exercise (Landy & Farr, 1983; Schwab, Heneman, & DeCotiss, 1975). In fact, Landy and Farr argued that since the variance accounted for in psychometric quality by rating formats was as low as 4%, a moratorium on rating format research should be followed.

To some however, the initial research question was still compelling: Does format make a difference relative to rating errors or the reliability and validity of the ratings? In 1997 and 2001, Borman and colleagues pursued the answer to this question by broadening the concept of rating "format" to include measurement models and raters' cognitive processes. They argued that rating formats were designed to first help raters organize their search for ratee behaviors, and then translate this observed behavior into data appropriate for making accurate evaluative decisions about a ratee's performance. The most effective rating formats, they argued, should support raters' cognitive processes in the sense that the format naturally leads raters to make these accurate evaluative judgments about observed ratee performance. The aforementioned formats (e.g. BARS, Behavior Observation Scales, etc), although effective at helping raters organize and accurately evaluate observed ratee performance, do not follow the same cognitive processes a rater is engaged in during the rating exercise. During an exercise of rating performance, a rater is requested to evaluate the rating target against each performance dimension independently, using the scale provided. Once each individual rating is complete, the rater is asked to mechanically combine the competency-level ratings into an overall rating, which in a selection interview setting, is typically used to determine which candidate progresses to the next step in the selection process. However, it is our contention here, and that of Borman et al. (1997, 2001), that these aforementioned rating formats do not adequately ensure the rater follows this process and instead allow for subjective preconceptions to easily impact the overall rating.

To help illustrate this, consider the following example. After speaking with the candidate for a few minutes, an interviewer decides there is something they do not like

about the candidate- they have a bad first impression. The impression may not be based on anything job relevant, but this may or may not matter to the interviewer- after all, they might be working closely with the person in the future, and they understandably want to be around someone they like. During and after the interview, the interviewer uses a well-developed interview guide with competency based BARS that are derived from valid job analysis data. After the interview interaction is complete, the intent is for the interviewer to evaluate each job-relevant competency independently and objectively, choosing the level of effectiveness the interviewee demonstrated during the interview. Once each rating is completed, the interviewer can average the ratings across competencies, and arrive at a numerical representation of the interviewee's performance. Because the interview guide was developed competently, high values should be predictive of excellent future job performance, whereas low values indicate the interviewee may not be qualified for the job. Of course, this is all readily apparent to the interviewer. Because the interviewer may (consciously or unconsciously) want to ensure this particular candidate does not progress further in the selection process, when completing the BARS, the interviewer can easily distort their ratings (e.g. severity) such that each competency effectiveness rating is below average, which consequently leads to a low overall interview performance rating. As a result, the candidate is likely to be dismissed from the selection process.

In this example, the interviewer readily manipulates the BARS rating, based on their job-irrelevant and subjective first impression. In the next section, we describe a rating format that may help prevent this type of intentional distortion or manipulation of

interview ratings, by making less obvious and automating the manner in which individual ratings are combined into an overall performance rating.

Paired Comparisons

In 1997, Borman and colleagues proposed that incorporating paired comparison judgments into a performance rating task might allow raters to utilize a more natural cognitive rating approach. Pairwise comparison ratings have been used in the study of preferences, attitudes, voting systems, social choice, public choice, and in psychology, where it is often referred to as paired comparison (Thurstone, 1927).

Borman suggested that one could simplify the performance-rating task by presenting a pair of behavioral statements and asking the rater to choose the statement that better characterizes the performance of the rating target. Such a rating task does not require the rater to make overly complex cognitive decisions, and instead asks that he or she compare the level of performance described by each of the behavioral statements, and choose the statement that is more like the observed performance. Borman and colleagues also suggested using a computerized adaptive testing (CAT) algorithm to choose and administer items would provide more information about the performance of each ratee than a nonadaptive measure, in which the same sequence of items would always be used. By using paired comparison rating formats, the rater's cognitive load is reduced such that the rater must only to focus on choosing the more accurate behavioral descriptor. In addition, by using an item response theory-based algorithm to administer the behavioral statements (i.e., presenting item pairs), more information can be obtained with less opportunity for a rater to distort ratings based on subjective preconceptions.

Development and Evaluation of CARS

As mentioned above, it was thought that using paired comparisons in the context of the performance rating process would make it easier for raters to produce a valid estimate a ratee's performance level. In 2001, Borman and colleagues suggested using an item response theory (IRT) –based computerized adaptive testing algorithm to make the paired-comparison ratings. By using IRT-based CAT technology to construct the paired comparisons measure and score responses, each successive rating judgment would provide a more precise estimate of a ratee's performance. More specifically, in the CARS rating process, pairs of behavioral statements would be presented one at a time, and a rater would be asked to choose the statement in each pair that is more descriptive of a ratee's performance. Each behavioral statement would be associated with the same performance dimension, and each paired comparison item would be composed of statements varying in levels of effectiveness. The choice of items and the scoring of responses would be carried out by an adaptive testing algorithm designed to provide the maximum amount of information with relatively few items, and rating would continue until a ratee's level of effectiveness could be estimated with sufficient precision. As demonstrated by Stark and Drasgow (1998), CAT results in more precise estimates of latent trait values (conceptualized as standard error of measurement) than nonadaptive tests nearly twice as long.

The relative improvement of CARS over a more conventional performance appraisal formats was explored further by Borman et al. (2001). In their study, the authors examined the comparative reliability, validity, and accuracy of the CARS format relative to GRS (graphical rating scales) and BARS. The authors chose to apply the

CARS format to the contextual performance domain (commonly also referred to as organizational citizenship behavior), an area that traditionally receives less empirical attention than task performance. Using videotaped vignettes of actors performing various scripts of contextual performance behaviors, Borman et al asked 114 business persons to rate performance using the CARS and either the BARS (behavioral anchored rating scale) *or* the GRS (graphic (the numerical) rating scale). Results showed 23%-37% lower standard error of measurements (SEM) for the CARS format, and indicated that ratings using this paired-comparison rating task produced significantly higher estimates of both validity estimates (d = .18) and Cronbach's accuracy coefficients (median effect size of .08).

Current Study

Interestingly, despite such favorable results, little further research concerning the CARS format has been published. One notable exception is the simulation study conducted by Schneider and colleagues (2003). These authors developed CARS to measure the entire managerial performance domain, including task and citizenship performance. Their work showed it is possible for a computer adaptive rating scale to measure facets of task performance- a limitation of the earlier CARS version. However, these authors tested the flexibility of CARS by simulating the actual rating exercise, and did not use actual performance data or ratings.

Accordingly, our notion at this point is to continue to explore this potentially useful rating format, and extend the research in a number of ways. First, we plan to follow similar steps as described by Borman et al. (2001) to develop a computer adaptive rating scale. However, we intend to focus on task performance, as operationalized by

Motowidlo and Burnett (1995). These authors provided a relatively concise task performance model- containing only four factors, thus making it an appropriate model to evaluate, at least initially, related to how CARS functions in the task performance domain. A more detailed description of these performance dimensions can be found in the next chapter. Second, we intend to evaluate the use of CARS in an interview evaluation setting. The intent is to extend the application of CARS, and provide HR professionals with options other than BARS when designing selection procedures such as structured interviews. As mentioned previously, BARS provide ample opportunity for subjective preconceptions to affect ratings, and appear to require a rater to follow a more complex cognitive process than a paired-comparison rating task. Finally, we will compare the criterion-related validity estimates of the paired-comparison computerized adaptive rating scales to the more popular behavioral anchored ratings scales (BARS) format.

Hypotheses

We believe that CARS will be a valid predictor of job performance in an employment interview setting, and that CARS will prove to be a superior rating format than BARS. To this end, we propose the following hypotheses.

Based on previous research trends, we expect interview outcomes produced by participants using CARS to be valid predictors of job performance, as operationalized by supervisor ratings of interviewee job performance. Due to the iterative and adaptive nature of CARS ratings, criterion related validity coefficients should be positive and significant.

H1: Performance ratings of interviewees produced by Computer Adaptive Rating Scales (CARS) will be positively and significantly correlated with the interviewee's supervisors' job performance ratings.

The second hypothesis is that the CARS ratings will be more predictive of job performance than CARS. We believe the greater precision attained with the CARS would result in a stronger relationship between interview evaluation ratings and supervisory performance ratings than what will be found with BARS.

H2: Computer Adaptive Rating Scales (CARS) will yield significantly larger predictive validity coefficients with supervisory performance ratings than Behavioral Anchored Ratings Scales (BARS).

Chapter 2: Method

In this section, we describe the performance domain targeted by CARS and BARS; discuss the steps taken to develop the CARS program; describe the development of the CARS and BARS scales; review the interviews used as stimuli; describe the supervisor job performance ratings; and then present the data collection methodology used.

Performance Domain

As mentioned previously, one of the goals of this study was to extend the study of CARS into the task performance domain. Previous research (e.g. Borman et al. 2001) examining CARS focused on contextual performance, or relied on simulated results to measure task performance (e.g. Schneider et al., 2003). In this study, task performance was defined using Motowidlo and Burnett's (1995) four dimension taxonomy-leadership, teamwork, planning and organizing, and drive. These dimensions of management effectiveness were based on results of numerous job analyses and studies of management positions across a number of different jobs and organizations (e.g. Motowidlo et al., 1992).

The task performance dimensions used in this study are as follows. ***Leadership*** refers to seeking opportunities for leadership, directing and guiding others toward the accomplishment of tasks by motivating and assessing their performance and/or behavior, persuading others to accept own ideas and exhibiting confidence in those ideas, and

taking initiative, taking charge. ***Teamwork*** refers to emphasizing and showing concern for group interests, cooperating with others and working to form harmonious work groups, prioritizing group interests above individual interests, helping and listening to others, and showing consideration for the needs and feelings of others. ***Planning and Organizing*** refers to one's ability to adopt a methodical and systematic approach for solving all aspects of a problem, giving specific attention to detail, generating and evaluating alternative solutions thoroughly, anticipating obstacles and developing plans to meet them, and setting appropriate priorities. Finally, ***Drive*** is defined as showing concern for task accomplishment, persisting to solve problems and overcome obstacles to task accomplishment, doing extra work and focusing high energy levels to solve problems or meet difficult deadlines, and volunteering to handle assignments or problems outside own area of responsibility.

The current study intentionally did not make any changes to this taxonomy to ensure comparisons could be made between interview ratings collected here and job performance ratings produced by supervisors of the interviewees, which were the same as those used in Motowidlo and Burnett's research studies.

Scale development

The process to develop CARS for measuring interviewee performance in this study involved two steps. First, the computer platform used to administer the test was created. Second, behavioral statements were developed to populate the platform.

Building the CARS program. To build the computer platform for this study, the author began with a version of the computer program developed by Stark and Drasgow (1998). In the original program, behavioral statements were selected by estimates of the

interviewee's performance, as generated by the algorithm and the rater's choices for previous statement pairs. The platform begins by assuming the interviewee has an average level of effectiveness in the current dimension. A pair of statements, one lying above and one lying below that trait estimate, is selected subject to symmetry and information constraints. The rater is initially asked to choose the statement of the pair that is more descriptive of the interviewee. This response is then scored using Bayes modal estimation, and a new effectiveness estimate is formed. A search process for additional statement pairs then begins, ending when the desired number of items has been administered or the algorithm can no longer find pairs that provide at least 50% of the theoretical maximum item information at the interviewee's most recent effectiveness level. Early termination of the search process tends to occur only when the ratee's performance is either very high or very low, given a statement pool of reasonable size (Borman et al., 2001). Previous applications of the CARS limited the maximum number of pairs presented for a single ratee and dimension in an effort to minimize rating time (Borman et al., 2001; Schneider et al., 2003). To this end, a maximum limit of 8 pairs was imposed in the current CARS platform.

The original CARS programs were written using the Visual Basic programming language (Borman et al., 2001; Schneider et al., 2003), but the current study used a web-based PHP Hypertext Preprocessor (PHP) programming language to administer Stark and Drasgow's (1998) paired-comparison algorithm in an online format. The PHP-based program allows for more flexible administration of stimuli (e.g. web-based video taped interviews) and automated remote data collection. The development and quality-

assurance testing for the PHP based CARS program took approximately 5 months to complete.

Development of Behavioral Statements. Behavioral statements were generated to reflect the example behaviors for each of the four managerial effectiveness dimensions at varying levels of effectiveness. To produce these statements, we began by using subject matter experts (five advanced Industrial/Organizational graduate students), who were asked to provide as many statements as possible reflecting the varying levels of effectiveness for each of the four performance dimensions. The author then compiled all the generated statements and edited them for content, grammar, and writing style. Approximately four hundred statements (approximately one hundred per dimension) were initially produced.

At this point, the author along with one other SME categorized each statement into one of the four performance domains, and simultaneously arrived at a consensus rating for the effectiveness level of each statement on a 7-point scale (1 = very ineffective; 7 = very effective). Statements which could not be clearly categorized into a single dimension, or for which a consensus could not be reached were either discarded or rewritten. This process took approximately two weeks to complete, and resulted in one hundred seventy-eight finalized statements, with approximately 45 statements per dimension. Examples of these behavioral statements are presented in the appendix.

BARS development. The BARS used in this study was identical to those used by Motowidlo and Burnett (1995). They were anchored at the high, moderate, and low effectiveness levels, with approximately 10 behavioral statements across the three

anchors. To aid in data collection and to control any effect paper-pencil rating may have, these BARS were uploaded into the same web-app as the PHP CARS scale.

Videotaped Interviews

To compare BARS and CARS rating outcomes in an employment interview setting, we first required access to structured employment interviews. To this end, we obtained a total of 45 videotaped interviews used in a series of studies conducted by Motowidlo and Burnett (Motowidlo & Burnett, 1995; Burnett & Motowidlo, 1998). Each of the 45 videos depicted managers at telecommunication and utility companies across the southeast United States answering four behavioral-based interview questions. The managers were asked to answer the questions as if they were applying for their present positions, and therefore were instructed to only provide examples of actual behavior in past situations that occurred prior to their current jobs.

All interviews were originally conducted by the same female interviewer, and were recorded on VHS videotape. The video camera was placed behind and over the shoulder of the interviewer. Therefore, the interviewer was not visible to the viewer. Interviewees were seated at a table during the entire interview, with only their torso, head, and arms visible. Interviews ranged from 8 to 36 minutes with the average interview lasting 21 minutes.

To prepare the VHS tapes for this study, each interview was digitally encoded and edited to eliminate background noise (likely due to aging VHS tapes) and other study-irrelevant material. Each video was examined carefully for clarity of sound and picture. Each of the videos was then uploaded to a server such that they could be viewed and rated remotely by study participants.

Job Performance Ratings

To compare the validity of the ratings produced by the BARS and CARS conditions, we also required estimates of each of the interviewees' job performance, to serve as criteria for our criterion-related validity analyses. For this purpose, we obtained job performance ratings produced by supervisors of the participating interviewees in the Burnett and Motowidlo studies. Each supervisor had rated his or her direct reports on the four performance dimensions- leadership, teamwork, planning and organizing, and drive, using the BARS scale described earlier. A composite job performance rating was calculated by summing the individual dimension scores, and represented the overall job performance across the task performance domain for each interviewee.

Two of the original supervisor ratings were missing from the dataset provided by Burnett and Motowidlo and thus the corresponding interviews were not usable in the current study. As a result, a total of 43 interviewee videos and associated supervisor performance ratings were available for use in this study. Correlations between performance dimensions ranged from .36 to .56, with a mean of .46.

Data collection

The unit of analysis for this study is the number of interviews (43). Three raters were recruited to observe and rate each videotaped interview. Raters all had at least a master's degree, and three years of work experience, during which their responsibilities included interviewing candidates for a role in their organization.

Two of the raters were assigned to the CARS condition, and the third was assigned to complete the BARS condition. Because the original Burnett and Motowidlo data included BARS interviewee ratings produced by the female interviewer, it was

decided to include these initial ratings in the current study and have ratings from two independent sources in each of the study conditions (BARS and CARS).

Although more raters would be desirable to ensure reliability of the interview ratings, this would have caused significant data collection challenges in the current study. Given the 43 videos averaged approximately 21 minutes in duration, rater fatigue would likely have led to significant data reliability issues. In fact, a previous attempt by the author to obtain similar ratings was severely limited by these issues, despite the fact that the videos used in the previous study were only between 5-10 minutes long (Schmidt, 2007).

Prior to data collection, the author contacted each rater and provided a 30-minute orientation to the web-app and rater-error training session. Each rater was given a unique username and password to the web-app, and was instructed to log in, view at least one video in its entirety, take notes as necessary, and complete the rating exercise. The web-app was designed such that each rater could stop at any point (provided they were between interviews), and return at another time to continue. The web-app kept track of their progress, and upon their return, picked up with the next video in queue, thus allowing raters to complete the extensive rating task at their own pace and as their schedules allowed.

With an average video length of 21 minutes, the total time needed to complete either BARS or CARS rating after viewing a single video was approximately 30 minutes. With 43 total videos, this involved approximately 22 hours of commitment, per rater. In total, it took approximately 60 days for all three raters to complete their ratings.

Similar to the supervisor ratings, dimension-level ratings produced in the BARS and CARS conditions were summed into a composite variable, representing an overall rating of interviewee effectiveness.

In addition to collecting interview ratings through CARS or BARS, we asked each rater to provide their perceived level of confidence with the rating task after completing the rating for each of the 43 videos. Specifically, we asked each rater to answer the following question: "How confident are you that your ratings were an accurate representation of this candidate's effectiveness during the interview?" Ratings were made on a 7-point scale (1= Not confident at all; 7= Extremely confident).

Chapter 3: Results

A total of 43 videos were rated in both the BARS and CARS conditions. In the subsequent analyses, the 'Total' variable refers to the composite of all dimension scores across BARS, CARS, and supervisor ratings, and should be considered the main variable of interest for this study. Again, the unit of analysis for all analyses is the interviewee (N=43).

CARS and BARS ratings produced by the two raters within each condition were averaged together for all subsequent analyses. Prior to doing so, however, ratings within each condition were reviewed by the author for any major discrepancy in ratings. Agreement between the two raters was assessed using Shrout and Fleiss's (1979) intraclass correlations (Case 2, 1-rater). In the BARS condition, the intraclass correlation was 0.59, while in CARS the computed intraclass correlation was 0.75. Descriptive statistics for the study variables can be found in Table 1.

Table 1 – Descriptives for Overall and Dimension-level ratings across conditions

	BARS		CARS		Supervisor Ratings	
Variable	Mean	SD	Mean	SD	Mean	SD
1. Total	16.42	3.87	17.74	3.85	21.51	3.36
2. Leadership	3.77	1.26	4.35	1.15	4.79	1.26
3. Teamwork	4.13	1.23	4.66	1.07	5.77	.84
4. Planning & Org.	4.28	1.06	4.20	1.08	5.40	1.05
5. Drive	4.24	1.05	4.53	1.06	5.56	1.14

Note: Dimension-level scales range 1-7; Total scale range 4-28

Hypothesis 1 concerned the relationship between CARS and supervisor ratings. To test Hypothesis 1, criterion-related validity coefficients were computed for each

dimension as well as for the total overall score variable. In addition, one-way ANOVAs were computed for each of the four dimensions and total scores.

Examining each dimension individually provides some insight into differences in ratings. Beginning with Leadership, a one-way ANOVA revealed significant differences between groups ($F(2, 126) = 7.44$, $p < .01$, and post-hoc tests show that average BARS ratings ($M = 3.77$, $SD = 1.26$) were significantly lower than both CARS ($M = 4.35$, $SD = 1.15$) and supervisor ratings ($M = 4.79$, $SD = 1.26$), but no difference was found between CARS and supervisor ratings. Table 2 depicts correlations between rating condition and supervisor performance ratings.

Table 2 – Correlations between BARS, CARS, and Supervisor Performance Ratings for Leadership Dimension

Variable	1	2	3
BARS	--		
CARS	0.30*	--	
Supervisor Rating	0.23	0.45**	--

Note: '' is significant p < .05; '**' is significant p <.01*

In the Teamwork dimension, ANOVA results were significant ($F(2, 126) = 29.87$, $p <.01$), with post hoc tests indicating that both BARS ($M = 4.13$, $SD = 1.23$) and CARS ($M = 4.66$, $SD = 1.06$) ratings were significantly lower than supervisor ratings ($M = 5.76$, $SD = 0.84$). Table 3 depicts correlations between the variables.

Table 3 – Correlations between BARS, CARS, and Supervisor Performance Ratings for Teamwork Dimension

Variable	1	2	3
BARS	--		
CARS	0.62**	--	
Supervisor Rating	0.12	0.26	--

*Note: '**' is significant p <.01*

In the Planning & Organization condition, ANOVA results again indicated a significant group difference ($F(2,126) = 17.06$, $p < .01$), and post hoc tests revealing that supervisor ratings ($M = 5.39$, $SD = 1.05$) were significantly higher than both BARS ($M = 4.27$, $SD = 1.06$) and CARS ($M = 4.19$, $SD = 1.07$) ratings. Table 4 depicts correlations between the study conditions and supervisor ratings.

Table 4 – Correlations between BARS, CARS, and Supervisor Performance Ratings for Planning & Organization Dimension

Variable	1	2	3
BARS	--		
CARS	0.46**	--	
Supervisor Rating	0.28	0.39*	--

Note: '' is significant p <.05; '**' is significant p <.01*

In the final dimension, Drive, ANOVA results also confirmed a significant group difference ($F(2, 126) = 17.50$, $p < .01$), with post hoc tests indicating that BARS ($M = 4.23$, $SD = 1.05$) and CARS ($M = 4.53$, $SD = 1.06$) were significantly lower than supervisor ratings ($M = 5.56$, $SD = 1.14$). Correlations between the variables can be found in Table 5.

Table 5 – Correlations between BARS, CARS, and Supervisor Performance Ratings for Drive Dimension

Variable	1	2	3
BARS	--		
CARS	0.44**	--	
Supervisor Rating	0.18	0.26	--

*Note: '**' is significant p <.01*

When looking at the composite BARS, CARS, and supervisor ratings, supervisor ratings ($M = 21.51$, $SD = 3.36$) were significantly greater than both BARS ($M = 16.42$, $SD = 3.87$) and CARS ($M = 17.74$, $SD = 3.85$) ($F(2,126) = 21.91$, $p < .01$). Correlations

between composite interview ratings by study condition and supervisor job performance ratings can be found in Table 6.

Table 6 – Correlations between BARS, CARS, and Supervisor Performance Ratings for Total Performance***

Variable	1	2	3
BARS	--		
CARS	0.59**	--	
Supervisor Rating	0.38*	0.36*	--

Note: '' is significant p <.05; '**' is significant p <.01; *** Total performance refers to a composite variable produced by summing dimension level ratings*

A review of the relationships between the rating conditions and supervisor job performance ratings suggest support for Hypothesis 1. In order to test Hypothesis 2, which predicted CARS ratings would be significantly more predictive of performance than BARS, we examined whether there were any differences between the validity coefficients of BARS-supervisor ratings and CARS- supervisor ratings. To do so, we tested for the equality of dependant correlations with one variable in common, first looking at the composite Total rating variables. There was no significant difference found between BARS-supervisor ratings and CARS-supervisor ratings validity coefficients ($t(40) = 0.15$, $p > .05$). In addition, results of relative weights analysis (Johnson, 2000) using rescaled relative weights (computed by dividing each relative weight by the R^2 in order to get a percentage of predicted criterion variance attributable to each predictor) show that BARS explained slightly more variance in performance than CARS (54.3% BARS versus 45.7% for CARS). Taken together, these results do not provide support for Hypothesis 2.

Whereas the second hypothesis was not supported, further analyses were conducted to examine whether correlations between dimensional ratings and supervisor

ratings across study conditions differed significantly. Additional tests for equality of dependant correlations with one variable in common indicated that there were no significant differences between BARS-supervisor ratings and CARS-supervisor ratings in any of the four performance dimensions. Results of these analyses can be found in Table 7.

Table 7 – Test for Equality of Dependant Correlations with one common variable

Dimension	BARS-Supervisor *r*	CARS-Supervisor *r*	*t*
Total*	0.38	0.36	0.15
Leadership	0.23	0.45	-1.31
Teamwork	0.12	0.26	-1.05
Planning & Organization	0.28	0.39	-0.74
Drive	0.18	0.26	-0.50

**Total performance refers to a composite variable produced by summing dimension level ratings*

As mentioned earlier, a one-item measure of raters' confidence that their ratings were an accurate representation of the interviewee's actual effectiveness was administered after each video and test condition was completed. Although no hypotheses were proposed around this measure, it could provide insight into users' sentiments toward BARS and CARS. For the CARS condition, confidence ratings were averaged across the two participants. An independent samples t-test indicated that the mean confidence rating in the CARS condition (M = 4.88, SD = 0.69) was significantly greater than the mean confidence rating in the BARS condition (M = 4.29, SD = .77) [$t(84) = -3.76$, $p < .01$].

Chapter 4: Discussion

This study sought to investigate the relationship of interview ratings produced using behavioral anchored rating scales (BARS) and computer adaptive rating scales (CARS). Existing research suggests ratings made using CARS are significantly more predictive of job performance ratings than those made using BARS. In addition, the research on CARS had been limited predominantly to the contextual performance domain, and focused primarily on performance appraisal processes. The current research sought to extend the research on CARS to the task performance domain and examine the potential of CARS as an alternative rating format in structured employment interviews.

Results of this study indicate that ratings produced by participants in the CARS condition are valid predictors of job performance, demonstrating clear support for the first hypothesis. Across the four performance dimensions, CARS was positively and significantly correlated with supervisor ratings of job performance in both the Leadership ($r = 0.45$) and Planning & Organization dimensions ($r = 0.39$). When the dimension level ratings were combined and composite rating and criteria variables were created, CARS was significantly correlated to job performance ($r = 0.36$). Interestingly, none of the dimension-level relationships were significant in the BARS condition, but when the dimension ratings were combined into a composite variable, the relationship between BARS rating and job performance was significant ($r = 0.38$).

Based on dimension-level relationships, it was anticipated that the pattern of CARS out-performing BARS would follow. A potential explanation for this finding could be attributed, at least partially, to low rater agreement in the BARS condition. More specifically, intraclass correlations were substantially greater in the CARS condition than in the BARS condition (0.75 versus 0.59, respectively), indicating that BARS ratings may not have predicted dimension-level performance as well as CARS ratings due to the reliability of the ratings. It is therefore possible that the significant validity coefficients produced by correlating the mechanical composite of dimension-level BARS ratings with supervisor ratings is a spurious finding.

Subsequent research could explore this possibility further by using more raters in each condition to ensure more stable estimates of rater agreement. In addition, raters in the BARS condition should be asked to provide an overall rating of interviewee performance, which could be examined alongside the composite variable used in the current research. Doing so may help shed additional light on the discrepancies between the dimension-level and composite-level validity estimates found in this study.

The remaining analyses focused on whether CARS provided any incremental level of prediction, above and beyond BARS. It was expected, based on previous research (e.g. Borman et al., 2001, Schneider et al., 2003), that the iterative and adaptive nature of the CARS rating format would lead to more valid estimates of job performance. However, results of the current study did not follow this pattern. In fact, CARS-supervisor rating and BARS-supervisor rating relationships were not significantly different from one another. Furthermore, a relative weights analysis found that BARS

accounted for slightly more variance in job performance than CARS (54.3% and 45.7%, respectively).

This study also contained a one-item measure of rater confidence. Raters in the CARS condition were significantly more confident than those in the BARS condition, although both groups were confident their ratings were representative of interviewee effectiveness (mean confidence was 4.88 for CARS and 4.28 for BARS, on a 7-point scale). Although no differences were found between the validity coefficients produced by the CARS and BARS conditions, it is interesting to note that raters felt more confident in the CARS condition.

Theoretical Implications

The results of this study indicate a need for further research. This study sought to extend CARS research into two new areas - the task performance domain and the employment interview, both of which may have had some impact on the null findings, and should be examined further.

Previous CARS research did show promising results when used in the task performance domain (e.g. Schneider et al., 2003), albeit with simulated data instead of a field or lab study. Clearly, the current study was unable to replicate this success, raising some questions around the effectiveness of CARS for task performance rating research. The conceptual model of contextual performance used in previous studies (e.g. Borman et al., 2001) was comprised of three dimensions, and was the result of much empirical examination. In short, it is fairly certain that model covers the entire conceptual domain of contextual performance with three dimensions, and has repeatedly been shown to be stable across jobs and organizations (e.g. Conway, 1996; Coleman & Borman, 2000). As

a result, participants in their study were likely familiar with the facets of contextual performance from their own personal experiences, and had a good sense of effective and ineffective performance levels across the three dimensions regardless of their own job and industry history. As such, it is likely that any behavioral examples of contextual performance in their study were easily categorized into one of the three performance dimensions.

Task performance dimensionality, however, is typically driven by the specific job, organization, and industry it is developed for (Conway, 1996; Coleman & Borman, 2000). Although the model of task performance used in this study focused on managerial performance, and was developed and successfully tested across a number of previous studies (e.g. Motowidlo & Burnett, 1995; Burnett & Motowidlo 1998), it is possible the model may not have the same extent of conceptual domain coverage as the aforementioned contextual performance model. In addition, it is likely that raters in this study were not as familiar with the task performance model based on personal experiences. As such, it is possible that it was relatively more difficult for raters in this study to observe, categorize, and rate behavior of such interviewees, undermining some of the benefits of CARS found in previous studies.

For example, Borman et al. (2001) argued that CARS results were more valid, reliable and accurate than other rating formats because it presented more behavioral statements that are more targeted toward the ratee. In other words, the authors argued that the adaptive nature of CARS allows for more precise ratings of a rating target. However, if an interviewee's responses to interview questions did not fall clearly within one of the four dimensions of the task performance domain, or if the participant was not very

familiar with the performance model, no amount of adaptive rating or precision can overcome them relying on biases or heuristics to arrive at their rating.

In addition, because task performance varies by job and industry, it may not have been a strong framework to test CARS validity given the design of the current study. Specifically, the criterion measure in this study was produced by the interviewee's supervisors- individuals who likely have a keen understanding of what constituted effective performance across each of the four dimensions. This level of understanding did not appear to extend to raters in this study. For teamwork, for example, average supervisor ratings in this dimension were considerably higher than those produced in both the CARS and BARS condition (5.76 versus 4.66 and 4.13, respectively). If supervisors do in fact have a more nuanced understanding of effective teamwork performance for their associates, it is possible that the same behaviors that participants in this study felt constituted a 3 or 4 were viewed as a 5 or 6 by the supervisors. As such, validity coefficients (and thus any potential differences between CARS and BARS relationships with job performance) may have been impacted due to performance standard differences between study participants and supervisors.

Future research should, however, continue to focus on extending CARS research into the task performance domain, but be careful when choosing the specific model. In addition, future researchers intent on studying task performance should ensure both raters of performance and supervisors are calibrated to have the same standards of performance.

As mentioned earlier, the current study also attempted to move beyond the performance appraisal process, and examine the validity of CARS as a rating format in the employment interview process. This may also have factored in to the null findings

found when testing the second hypothesis. A significant difference between these two human-resource processes is the length and type of exposure to ratee behavior. In the performance appraisal process, raters typically observe ratees over a considerably longer amount of time (e.g. an employee's entire tenure as a direct report to a manager), and are exposed to a significantly larger sample and variety of ratee behavior. When the actual performance appraisal process eventually takes place, the rater likely has a larger pool of data to draw from when choosing a particular behavioral statement within the CARS program. In the employment interview, the rater is exposed to a limited sample of behavior. Within the CARS rating task therefore, the rater may not have enough data to draw upon in order to choose the valid behavioral statement. Previous research conducted by Borman et al. (2001) focused on the performance appraisal process, and found much more promising results for CARS. While their research also used videotaped stimuli, the vignettes presented to their raters were purposefully scripted to contain a rich enough pool of data for raters to produce ratings without having to rely on heuristics or biases.

Another significant difference between performance appraisal and employment interview processes concerns ratee motivation. Despite the current study using 'for research purposes only' interviews as stimuli, the likely experienced and talented interviewees in the videos were probably sharing examples of their peak performance, or describing situations which highlighted their strengths. Even if a response to a behavioral interview question contains less desirable behavior, experienced, talented, or motivated interviewees are likely to distort, overlook, or rationalize any example of poor performance to ensure they appear competent to the interviewer. As a result, raters in both conditions in the current study were likely basing their ratings on peak performance

data. This may have veiled any variation in true performance, and undermined any true variance in ratings due to rating format differences. On the other hand, in performance appraisal settings, managers are exposed to both peak and typical performance levels of their direct reports. The opportunity to observe a larger variation in performance likely has some impact on the raters' ability to leverage the full potential of a rating scale.

In general, although current results indicate CARS is a viable alternative to BARS in predicting future job performance based on interviewee behavior, further research is needed to conclude that they are, in fact, differentially predictive.

Applied Implications

A major goal of this study was to examine whether an applied practitioner should consider using CARS in an employment interview setting. As discussed previously, organizational psychologists are usually interested in any process improvement or format manipulation that could yield a greater return on their investment. CARS appear to be one such option, with previous research demonstrating promising validity, reliability, and accuracy estimates. The criterion-related validity analysis in the current study supports this notion as well.

However, had the pattern of results found in earlier studies of CARS performance ratings confirmed and demonstrated an increase in validity over other rating formats, a logical next step would have been to extend this research to a true field sample and to build CARS for hiring managers to use on real job applicants. Unfortunately, the uncertain results found in this study likely will prevent most practitioners from doing that. At this time, it would be prudent to call for more research in this area, specifically

around using CARS to predict task performance in the employment interview, in a field setting, before committing more resources to this end.

Another consideration for applied practitioners concerns the development time and resources needed for CARS. Whereas BARS are somewhat quickly and easily developed, preparing an item bank of behavioral statements for a new CARS task performance domain is substantially more time-consuming.

Finally, it must be noted that the rating task itself takes significantly longer in the CARS condition than in other rating formats. In this study, participants assigned to the CARS condition noted that the format was significantly more tedious to complete than other formats they had experienced. As such, it may be especially challenging for practitioners to convince organizational decision-makers to pursue the development of CARS in their organizations, especially when it would be used as only one part of a multi-hurdle selection process.

Limitations

In addition to those discussed above, there are a number of other limitations that could be contributing to the uncertain findings of this study. First, although the current research used real interviewees and performance data obtained from managers and supervisors in actual organizations, the rating exercise was still conducted in a lab setting, with stimuli presented via computer. Particularly when a lab study's data collection is lengthy, repetitive, and computerized, participants have more freedom to multi-task. Similarly, participants were not required to interact with the rating target (as they would have been in a field study where they would be seated in the same room as the interviewee), which may have led to even less consistent attention. Of course, in this

study, both the CARS and BARS conditions suffered equally from these limitations, but any potential differential effect should be examined further.

As described earlier, ratings produced by two individual raters were averaged for use in subsequent analyses. Interrater reliability estimates for each condition were satisfactory, but this issue raises another limitation for the study. More specifically, it would have been desirable to have more raters in each condition, which likely would have enabled us to produce more stable reliability estimates to propose hypotheses around format reliability.

In addition, another limitation is the psychometric properties of the criteria used in the study. Supervisors completed their ratings using a BARS scale, introducing a potential confound into the study. Subsequent research should consider using multiple rating formats to collect criteria data to control for any such effects. In addition, there was no measure of validity or reliability of the criteria, and therefore no way to correct or control for these issues. Future studies should also examine the potential to include more objective criteria to determine the impact of these shortcomings.

The final limitation is the sample of interviewees. As noted previously, interviewees were instructed to act as though they were applying for their current jobs, despite already having been selected into the organization, presumably passing some unknown selection process to do so. This restricted range of interviewees may have played a role in the results found here. Supervisor ratings were higher than both BARS and CARS across all four performance dimensions, which may either indicate a leniency bias on behalf of supervisors or reflect a restricted range that was not corrected for.

Summary

In closing, this research sought to extend the body of psychometric research of CARS while investigating its potential for use in the employment interview. Using previously recorded interviews and supervisor evaluations, and constructing a new task-performance based CARS, these notions were tested. Mixed evidence was found to support the use of CARS in the employment interview. With regard to validity, CARS ratings did predict job performance more effectively than with BARS on individual dimensions, although not when dimensions were combined into a composite variable. Because CARS is an extremely time- and resource-intensive measure to build and administer, these results cast some doubt on the utility and potential organizational acceptance of CARS as a viable alternative. Given the limitations of the current study, the book is not necessarily closed on using CARS in this manner. These findings, however, do suggest CARS should be further evaluated in the interview context.

References

Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. Applied Psychological Measurement, 17, 253-276

Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. Personnel Psychology, 35(2), 281-322.

Barnes, C. M., & Morgeson, F. P. (2007). Typical performance, maximal performance, and performance variability: Expanding our understanding of how organizations value performance. *Human Performance, 20*(3), 259-274.

Bendig, A. W. (1954). Reliability of short rating scales and the heterogeneity of the rated stimuli. *Journal of Applied Psychology, 38*(3), 167-170.

Blanz, F., & Ghiselli, E. E. (1972). The mixed standard scale: A new rating system. *Personnel Psychology, 25*(2), 185-199.

Borman, W. C. (1986). Behavior-based rating scales. In R. A. Berk, & R. A. Berk (Eds.), *Performance assessment: Methods & applications.* (pp. 100-120). Baltimore, MD US: Johns Hopkins University Press.

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology, 86*(5), 965-973.

Burnett, J. R., Fan, C., Motowidlo, S. J., & Degroot, T. (1998). Interview notes and validity. *Personnel Psychology, 51*(2), 375-396.

Burnett, J. R., & Motowidlo, S. J. (1998). Relations between different sources of information in the structured selection interview. *Personnel Psychology, 51*(4), 963-983.

Coleman, V.I and Borman, W.C (2000) Investigating the underlying structure of the citizenship performance domain. *Human Resource Management Review, 10,* 25-44

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*(5), 565-579.

Conway, J. M. (1996) Additional construct validity evidence for the task-contextual performance distinction. *Human Performance, 9,* 309-329

Coombs, C.H. (1950). Psychological scaling without a unit of measurement. *Psychological Review, 57,* 145-158

Cronshaw, S. F., & Wiesner, W. H. (1989). The validity of the employment interview: Models for research and practice. In R. W. Eder, G. R. Ferris, R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice.* (pp. 269-281). Thousand Oaks, CA US: Sage Publications, Inc.

Dipboye, R. L. (1992) *Selection interviews: Process perspectives.* Cincinnati, OH: South-Western.

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*(1), 35-49.

Gigerenzer, Gerd (200&). *Gut Feelings: The Intelligence of the Unconscious.* New York: The Viking Press

Gladwell, Malcolm (2005). *Blink: The Power of Thinking Without Thinking*. Boston: Little, Brown

Harris, M. M. (1989). Reconsidering the employment interview: A review of recent literature and suggestions for future research. *Personnel Psychology, 42*(4), 691-726.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology. 1,* 333-342

Huffcutt, A. I., & Arthur, W. (1994). Hunter and hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79*(2), 184-190.

Johnson, J.W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research, 35,*1-19.

Judge, T. A., Higgins, C. A., & Cable, D. M. (2000). The employment interview: A review of recent research and recommendations for future research. *Human Resource Management Review, 10*(4), 383-406.

Kane, J. S. (1986). Performance distribution assessment. In R. A. Berk, & R. A. Berk (Eds.), *Performance assessment: Methods & applications.* (pp. 237-273). Baltimore, MD US: Johns Hopkins University Press.

Krajewski, H. T., Goffin, R. D., McCarthy, J. M., Rothstein, M. G., & Johnston, N. (2006). Comparing the validity of structured interviews for managerial-level employees: Should we look to the past or focus on the future? *Journal of Occupational and Organizational Psychology, 79*(3), 411-432.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107.

Landy, F. J., Farr, J. L., & Jacobs, R. R. (1982). Utility concepts in performance measurement. *Organizational Behavior & Human Performance, 30*(1), 15-40.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 65*(4), 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/Psychologie Canadienne, 40*(1), 56-67.

Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. *Personnel Psychology, 30*(2), 255-268.

Madden, J. M., & Bourdon, R. D. (1964). Effects of variations in rating scale format on judgment. *Journal of Applied Psychology, 48*(3), 147-151.

Maurer, S. D. (1997). The potential of the situational interview: Existing research and unresolved issues. *Human Resource Management Review, 7*(2), 185-201.

Mayfield, E. C. (1965). Inside the interview. *PsycCRITIQUES, 10*(8), 365-366.

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*(4), 599-616.

Motowidlo, S. J., & Burnett, J. R. (1995). Aural and visual sources of validity in structured employment interviews. *Organizational Behavior and Human Decision Processes, 61*(3), 239-249.

Motowidlo, S. J., Carter, G. W., Dunnette, M. D., Tippins, N., Werner, S., Burnett, J. R., et al. (1992). Studies of the structured behavioral interview. *Journal of Applied Psychology, 77*(5), 571-587.

Pingitore, R., Dugoni, B. L., Tindale, R. S., & Spring, B. (1994). Bias against overweight job applicants in a simulated employment interview. *Journal of Applied Psychology, 79*(6), 909-917.

Posthuma, R. A., Morgeson, F. P., & Campion, M. A. (2002). Beyond employment interview validity: A comprehensive narrative review of recent research and trends over time. *Personnel Psychology, 55*(1), 1-81.

Roberts, J.S. & Laughlin, J.E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20,* 231-255.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262-274.

Schmidt, G.F. (2007). The Effect of Thin-Slicing on Structured Interview Decisions. Unpublished master's thesis, University of South Florida, Tampa, FL, Department of Psychology.

Schmitt. N. (1976). Social and situational determinants of interview decisions: Implications for the employment interview. *Personnel Psychology, 29,* 79-101

Schneider, R. J., Goff, M., Anderson, S., & Borman, W. C. (2003). Computerized adaptive rating scales for measuring managerial performance. *International Journal of Selection and Assessment, 11*(2), 237-246.

Schwab, D. P., Heneman, H. G., & DeCotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology, 28*(4), 549-562.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*(2), 149-155.

Stark, S., & Drasgow, F. (1998). Application of an IRT Ideal Point Model to Computer Adaptive Assessment of Job Performance. *Paper Presented at SIOP Conference, 1998*

Stark, S., & Drasgow, F. (2002). An EM approach to parameter estimation for the Zinnes and Griggs paired comparison IRT model. *Applied Psychological Measurement, 26*(2), 208-227.

Taylor, E. K., & Wherry, R. J. (1951). A study of leniency in two rating systems. *Personnel Psychology, 4*, 39-47.

Taylor, P. J., & Small, B. (2002). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and past behaviour employment interview questions. *Journal of Occupational and Organizational Psychology, 75*(3), 277-294.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273-286.

Tziner, A., & Kopelman, R. E. (2002). Is there a preferred performance rating format?: A non-psychometric perspective. *Applied Psychology: An International Review, 51*(3), 479-503.

Ulrich, L., & Trumbo, D. (1965). The selection interview since 1949. *Psychological Bulletin, 63*(2), 100-116.

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

Yun, G. J., Donahue, L. M., Dudley, N. M., & McFarland, L. A. (2005). Rater personality, rating format, and social context: Implications for performance appraisal ratings. *International Journal of Selection and Assessment, 13*(2), 97-107.

Zinnes, J. L., & Griggs, R.A. (1974). Probabilistic, Multidimensional Unfolding Analysis. *Psychometrika*, 39, 327-350

Appendices

Appendix A – Examples of CARS Behavioral Statements

Examples of Behavioral Statements from the Computerized Adaptive Rating Scales

Behavioral Statement	Dimension	Effectiveness Level
Conveys specific, observable and/or measurable expectations for performance from others	Leadership	Effective
Readily offers help or assistance to others, even when facing a heavy workload	Teamwork	Effective
Seeks and gathers as much relevant information as possible for solving all aspects of a problem	Planning & Organization	Very effective
Sets inappropriate timeframes for achieving goals (e.g. unclear, unrealistic, etc.).	Drive	Very ineffective