Measuring the Power of Learning.®

Research Report
ETS RR–17-28

Exploring Methods for Developing Behaviorally Anchored Rating Scales for Evaluating Structured Interview Performance

Harrison J. Kell

Michelle P. Martin-Raugh

Lauren M. Carney

Patricia A. Inglese

Lei Chen

Gary Feng

December 2017

Discover this journal online at
Wiley Online Library
wileyonlinelibrary.com

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Senior Research Scientist

Heather Buzick
Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Research Director

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Distinguished Presidential Appointee

Anastassia Loukina
Research Scientist

John Mazzeo
Distinguished Presidential Appointee

Donald Powers
Principal Research Scientist

Gautam Puhan
Principal Psychometrician

John Sabatini
Managing Principal Research Scientist

Elizabeth Stone
Research Scientist

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Gontz
Senior Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

RESEARCH REPORT

Exploring Methods for Developing Behaviorally Anchored Rating Scales for Evaluating Structured Interview Performance

Harrison J. Kell, Michelle P. Martin-Raugh, Lauren M. Carney, Patricia A. Inglese, Lei Chen, & Gary Feng

Educational Testing Service, Princeton, NJ

Behaviorally anchored rating scales (BARS) are an essential component of structured interviews. Use of BARS to evaluate interviewees' performance is associated with greater predictive validity and reliability and less bias. BARS are time-consuming and expensive to construct, however. This report explores the feasibility of gathering participants' responses to structured interview questions through an online crowdsourcing platform and using those responses to develop BARS. We describe the development of 12 structured interview questions to assess four applied social skills, elicitation of responses to these questions in the form of critical incidents from 68 respondents, and the creation of BARS from these critical incidents. Results indicate online participants are able to produce responses of sufficient quality to generate BARS for evaluating structured interview performance. We conclude by discussing limitations to this approach and future directions for research and practice.

Keywords Amazon Mechanical Turk; behaviorally anchored rating scales; crowdsourcing; employment interviews; performance appraisal; social skills; structured interviews

doi:10.1002/ets2.12152

Employment interviews are one of the most popular means of selecting personnel (Levashina, Hartwell, Morgeson, & Campion, 2014; McDaniel, Whetzel, Schmidt, & Maurer, 1994). Structured interviews that present all interviewees with the same standardized questions have higher validities for predicting job performance than unstructured interviews that vary the questions posed to each interviewee (Schmidt & Hunter, 1998). Standardized, anchored scales to rate responses to each question comprise one element of interview structure (Campion, Palmer, & Campion, 1997). Use of behaviorally anchored rating scales (BARS) tends to increase the reliability and predictive validity of structured interview scores (Taylor & Small, 2002) and may decrease bias against protected groups (Reilly, Bocketti, Maser, & Wennet, 2006). Unfortunately, the amount of time and effort required to create BARS is often excessive (Landy & Farr, 1980; Landy, Farr, Saal, & Freytag, 1976; Maurer, 1997).

The tension between quality and practicality has led some to question whether developing BARS is worth the benefits they demonstrate (Borman & Dunnette, 1975; Landy & Farr, 1980) and also to investigate means of reducing the time and labor necessary to construct them (Champion, Green, & Sauser, 1988; Green, Sauser, Fagg, & Champion, 1981). The purpose of this report is to explore a new approach that may reduce the workload involved in creating BARS: online crowdsourcing. Although crowdsourcing platforms such as Amazon Mechanical Turk (AMT, 2016) and CrowdFlower have been recommended as a viable source of data by some (Buhrmester, Kwang, & Gosling, 2011; Mason & Suri, 2012), the quality of the responses generated by online respondents has also been questioned (Hamby & Taylor, 2016). Our investigation is guided by a question, not a hypothesis: Will online participants produce critical incidents of sufficient quality to develop BARS from them? We are operating in a discovery-oriented mode for this study, simply looking to explore the possibilities of this approach (McCall & Bobko, 1990; Tukey, 1969). If we find promising results, they will have to be replicated, confirmed, and justified (Tukey, 1980).

This report is organized into five sections. The first section presents an overview of structured employment interviews and four applied social skills constructs identified as frequently being measured by structured interviews. The second section reviews BARS in depth, including the traditional procedure used to develop them and various efforts that have

Corresponding author: H. Kell, E-mail: hkell@ets.org

been made to shorten this procedure. The third section provides background information about the specific conditions of the current study. The fourth section is Method and Results. The final section is Discussion, which includes an exploration of the study's limitations and sketches of future directions.

Employment Interviews

The employment interview is a selection tool designed to predict job performance using applicants' oral responses to inquiries (McDaniel et al., 1994); it was introduced into the professional literature by Walter Dill Scott in 1915 (Salgado, 2001). Long-standing narrative review and meta-analytic evidence (e.g., Mayfield, 1964; Ulrich & Trumbo, 1965; Wagner, 1949) have consistently supported the use of structured interviews, in which questions and scoring guidelines are standardized across applicants, over unstructured interviews that allow questions to vary across applicants and lack a standardized scoring procedure. In addition to demonstrating superior psychometric characteristics, structured interviews also tend to yield relatively small ethnic group differences in comparison to other predictors (Huffcutt & Roth, 1998) and generally positive applicant reactions (Hausknecht, Day, & Thomas, 2004).

The types of questions used in structured interviews come in two major varieties: situational and past behavior. Situational questions (SQs; Latham, Saari, Pursell, & Campion, 1980) present people with hypothetical work situations and ask them to describe how they would respond to those situations. Past behavior questions (PBQs; Janz, 1982; Motowidlo et al., 1992) present people with work situations they likely experienced in the past and ask them to describe how they responded to those situations previously. PBQs operate under the assumption that past behavior is one of the best predictors of future behavior (the behavioral consistency principle; Schmitt & Ostroff, 1986; Wernimont & Campbell, 1968). Research has shown that PBQs generally yield higher validity coefficients than SQs (Taylor & Small, 2002). Further, although findings are not entirely consistent, PBQ scores tend to be more strongly related to personality and job experience, whereas SQ scores tend to be more strongly related to cognitive ability or, perhaps, job knowledge (Levashina et al., 2014).

Psychological Constructs Assessed by Interviews

Huffcutt, Conway, Roth, and Stone's (2001) meta-analysis pointed out that often more attention is given to the psychometric and structural characteristics of interviews than the constructs they assess. They identified seven major construct categories frequently assessed: applied social skills, cognitive abilities, interests and preferences, knowledge and skills, organizational fit, personality traits, and physical attributes. We chose to focus on applied social skills, given that 83% of the U.S. labor force currently works in the service sector (Lee & Mather, 2008), and alternative, less costly means are widely available to assess the other constructs identified. Applied social skills are defined as "the ability to function effectively in social situations" (Huffcutt et al., 2001, p. 904) and consist of four narrower psychological constructs assessed by interviews, which we review in turn.

Communication

Even technically skilled employees will be unsuccessful if they communicate poorly with others. It is not surprising that communication has been consistently identified as a key 21st-century skill for students and workers across a variety of frameworks (Carnevale, Smith, & Strohl, 2013; Casner-Lotto & Barrington, 2006; Finegold & Notabartolo, 2010; National Research Council, 2011). Huffcutt et al. (2001) conceptualized communication as "the ability to express (and receive) ideas and information clearly, accurately, and convincingly" (p. 900), and the National Research Council (2011) identified components of communication that are more "other" focused, such as listening and interpreting information. Thus, we define effective communication as not only expressing ideas effectively and accurately, but also as listening to the ideas and information provided by others. This skill has also been tied closely to teamwork because effective communication is needed to establish strong group bonds.

Communication skills are frequently assessed in interview settings and have been shown to be related to hireability. Huffcutt et al.'s (2001) meta-analysis identified 26 studies that measured communication in an interview setting. In some studies, communication skills were measured in a broad sense (i.e., rating the "communication skills" of the applicant; Dalessio & Silverhart, 1994; Landy, 1976), whereas other studies focused on specific aspects of communication. For example, appropriateness of an interviewee's verbal content and verbal fluency (Hollandsworth, Kazelskis, Stevens, &

Dressel, 1979), as well as pitch, pausing, and speech rates (DeGroot & Kluemper, 2007; DeGroot & Motowidlo, 1999), has been shown to impact hireability judgments. The degree of power in speech (i.e., the extent to which participants are clear and direct) has also repeatedly been shown to be related to perceived employability and competence in interview settings (End & Saunders, 2013; Gibbons, Busch, & Bradac, 1991; Parton, Siltanen, Hosman, & Langenderfer, 2002).

Leadership

Leadership has also been identified as a construct frequently evaluated in the employment interview; specific leadership skills assessed during interviews include motivating and coaching others, meeting demanding goals and developing people professionally, and delegating tasks to foster team dynamics and solve problems (Huffcutt et al., 2001). Hiring future leaders is critical to organizational success, because without solid leadership, businesses that would otherwise succeed may fail. Interviewers perceive candidates with leadership skills as adaptable, socially skilled, goal driven, and competitive and who have the potential to manage others, drive innovation, and provide a vision. Numerous leadership theories from various perspectives have been developed, including trait theories, behavioral theories that focus on the motivation and demonstrated skills of effective leaders, and interactional theories that posit that successful leadership is context specific and depends on the situation or relationship (Bass, 1992).

Attempts have been made to relate several of the multiplicity of perspectives on leadership to the Big Five personality traits. For example, across leadership overall (Judge, Bono, Ilies, & Gerhardt, 2002) and transformational leadership (Bono & Judge, 2004), a consistent picture has emerged via meta-analytic summaries: positive correlations between leadership evaluations and agreeableness, conscientiousness, extraversion, and openness to experience, but a negative correlation between leadership evaluations and neuroticism. This pattern of associations is also preserved for one aspect of transactional leadership—contingent reward—but is actually reversed for a second aspect of transactional leadership—active management by exception (Bono & Judge, 2004).

A meta-analysis using interactional theory that based effective leadership on the quality of the relationships between leaders and subordinates (e.g., mutual trust, respect; Schriesheim, Castro, & Cogliser, 1999) found that leadership defined in this dyadic, exchange-oriented way is associated with job satisfaction, performance, and organizational commitment (Gerstner & Day, 1997). This suggests that a quality relationship between a manager and direct reports can be instrumental in facilitating numerous positive work outcomes, including satisfaction, performance, retention, and commitment. These findings inform employment selection as well as training development because organizational training programs are one of the most common efforts aimed at improving employee productivity, managerial potential, and overall organizational effectiveness (Brungardt, 2011; Jain & Anjuman, 2013).

Negotiation and Persuasion

Negotiation and persuasion skills have been frequently identified as crucial to academic and workforce success (Carnevale et al., 2013; Casner-Lotto & Barrington, 2006; Finegold & Notabartolo, 2010; National Research Council, 2011). When employees are interacting with others, differences in perspectives, needs, and opinions often arise, and negotiation and persuasion skills are necessary to mitigate these differences. For the purposes of our assessment tool, we treated persuasion and negotiation skills as a single construct because of their similarity (cf. Huffcutt et al., 2001). Many social skills frameworks use a broad definition of negotiation and persuasion that focuses not only on influencing others and achieving individual goals but also maintaining harmony and trust, both within and between organizations (Carnevale, Gainer, & Meltzer, 1990). In order to achieve these results, effective negotiators must listen to others, take others' perspectives, and attempt to find solutions that are best for the group (National Research Council, 2011). These abilities are especially important for those in leadership positions, but are still relevant to virtually any employment position in which one interacts with others.

Attempts to measure negotiation skills occur across a variety of fields, such as K–12 education, higher education, and even cross-cultural research, using diverse methodologies such as self-report (Wang, MacCann, Zhuang, Liu, & Roberts, 2009), situational judgment tests (SJTs; Phillips, 1993; Wang et al., 2009), in-person role-plays (Page & Mukherjee, 2009), and game-based performance measures (Durlach, Wansbury, & Wilkinson, 2008). Because these skills have been repeatedly identified as critical for the workforce, they have also been assessed using interviews, scores on which have been shown to predict job performance (Bolanvich, 1994; Hoffman & Holden, 1993; Phillips, 1993).

For example, negotiation SJTs were shown to be effective in predicting supervisor ratings for a telephone collection position (Phillips, 1993), while a self-report measure of persuasion predicted applicant success in an engineering position (Bolanvich, 1994).

Teamwork and Interpersonal Skills

As most job types require individuals to work with coworkers, clients, and business partners, employers consider teamwork and interpersonal skills vital for most jobs today (Morgeson, Reider, & Campion, 2005), and as such, these skills are the most frequently evaluated social skills in the employment interview (Huffcutt et al., 2001). Today's business environment is characterized by fast-paced innovation and global competition, leading to the development of a team culture where workers are held accountable not only for simply performing a function, but also for contributing to broader organizational success. The "team" work environment makes collaboration a highly valued skill in the workplace because effective teams achieve goals more efficiently than individuals (Jordan, Ashkanasy, Härtel, & Hooper, 2002) and are more likely to produce innovative ideas (Taylor & Greve, 2006) that drive organizational success. Interpersonal skills reflect distinct abilities that facilitate social functioning (Gardner, 1993) and include the abilities to build rapport, work with a diverse group of colleagues, and influence others.

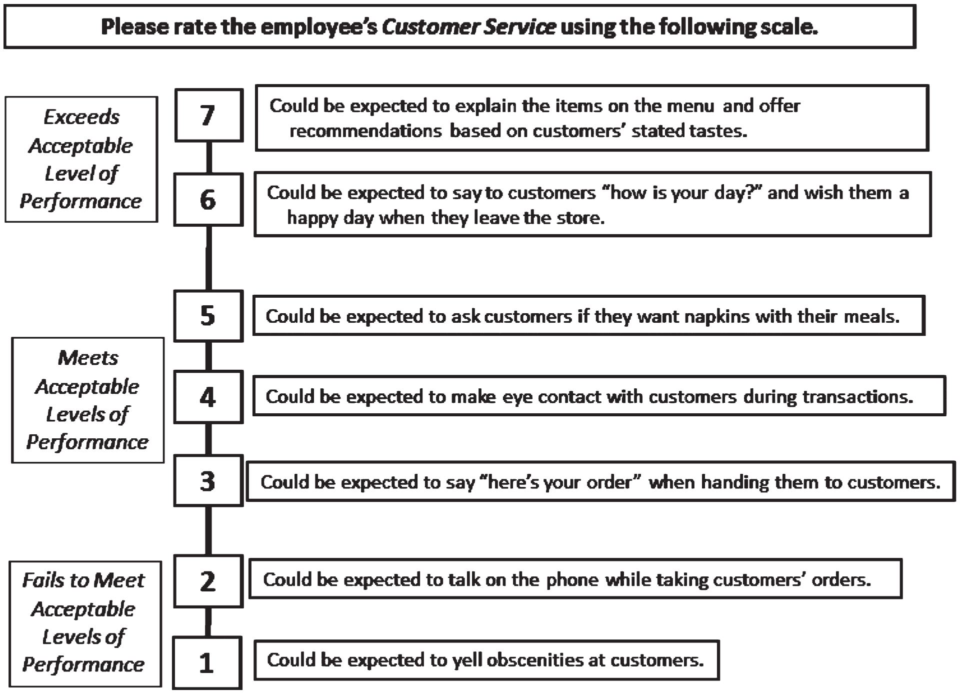
Behaviorally Anchored Rating Scales

BARS are generic term for scales that anchor an evaluative continuum with behavioral examples exemplifying performance at different levels of that continuum[1]; an example of BARS constitutes Figure 1. BARS were originally developed in the context of job performance appraisal (Smith & Kendall, 1963) as a means of reducing the influence of construct-irrelevant variance (Messick, 1989). Performance appraisal forms are often vague and ambiguous, failing to offer definitions of the performance dimensions being evaluated or of what performance at different levels (e.g., *above average*, *average*, *below average*) consists of (Schwab, Heneman, & DeCotiis, 1975). Consequently, raters' idiosyncratic interpretations of performance definitions and levels can result in different ratings, even when identical samples of behavior are observed. BARS aim to reduce the influence of these rater idiosyncrasies by defining performance in behavioral terms and offering concrete, specific examples of actions that exemplify performance at different levels (Smith & Kendall, 1963). BARS are an important part of what can be termed the "behavioral tradition" in industrial–organizational psychology (cf. Borman, 1986, 1991; Campbell, 1990, 2012; Dunnette, 1976; Motowidlo & Kell, 2013).

BARS are often derived from job analyses using the critical incident technique (Flanagan, 1954), which yields focused examples of workplace behavior provided by subject matter experts (SMEs) of the job(s) in question. These behavioral examples are analyzed for content similarities and grouped together according to those similarities to derive job performance dimensions that are, inherently, defined in terms of workers' actions. Edited versions of these critical incidents are used to anchor different levels of the evaluative continuum, providing specific examples of what various levels of performance constitute for each performance dimension. Not only do behavioral examples reduce the construct-irrelevance variance of raters' differing interpretations through provision of common reference points from which to base their judgments (Jacobs, Kafry, & Zedeck, 1980); ease of interpretation is facilitated because the terminology and content of the performance dimensions and behavioral anchors are derived from SMEs who are intimately familiar with the job of interest (Campbell, Dunnette, Arvey, & Hellervik, 1973).

Sometimes the dimensions being evaluated are framed in terms of job-relevant psychological attributes (e.g., cooperation, initiative; Borman & Dunnette, 1975) rather than performance dimensions (Kavanagh, 1971). In these circumstances, BARS can be used to restrict idiosyncratic interpretations as well, by focusing raters' attention on concrete behaviors that define those characteristics explicitly in the context of the specific job of interest, rather than abstractly or in "glittering generalities" (Guion, 2011, p. 461); raters are asked to perform less of an inferential leap by focusing on what ratees do at work—i.e., concrete actions—rather than who they are overall—i.e., generalized properties, or traits (cf. Anastasi, 1938; Cantor, 1990; Schwab et al., 1975), potentially leading to a decrease in illusory halo error (Borman, 1991).

BARS have proven to be popular far beyond the original job performance domain for which they were originally developed and have been used to assess classroom teamwork (Ohland et al., 2012), motivation (Landy & Guion, 1970), teaching effectiveness (Eley & Stecher, 1997; Hartsough, Perez, & Swain, 1998), morale (Motowidlo & Borman, 1977), and personality traits (Muck, Hell, & Höft, 2008). Due to their highly specific, highly work-relevant content, BARS have also



Please rate the employee's *Customer Service* using the following scale.

7 Could be expected to explain the items on the menu and offer recommendations based on customers' stated tastes.

6 Could be expected to say to customers "how is your day?" and wish them a happy day when they leave the store.

5 Could be expected to ask customers if they want napkins with their meals.

4 Could be expected to make eye contact with customers during transactions.

3 Could be expected to say "here's your order" when handing them to customers.

2 Could be expected to talk on the phone while taking customers' orders.

1 Could be expected to yell obscenities at customers.

Exceeds Acceptable Level of Performance
Meets Acceptable Levels of Performance
Fails to Meet Acceptable Levels of Performance

Figure 1 Example behaviorally anchored rating scale (after Smith & Kendall, 1963).

been suggested as being a viable basis for creating organizational feedback and training programs (Blood, 1974; Campbell et al., 1973; Hom, DeNisi, Kinicki, & Bannister, 1982). The following surveys the general procedures used to develop BARS.

General Procedures in the Development of Behaviorally Anchored Rating Scales

The fundamentals of creating BARS are relatively unchanged since the technique's inception, although many variations exist (see below). Consequently, BARS development has been described many times over the past 50 years, and the following draws liberally on both primary studies (e.g., Bernardin, LaShells, Smith, & Alvares, 1976; Campbell et al., 1973; Hedge, Borman, Bruskiewicz, & Bourne, 2004; Smith & Kendall, 1963) and reviews (e.g., Borman, 1986; Guion, 2011; Jacobs et al., 1980; Schwab et al., 1975).

Step 1

Critical incidents (Flanagan, 1954) are generated. Critical incidents typically depict specific, concrete, highly effective or highly ineffective workplace behaviors. A critical incident comprises the background situation that elicited the behavior, the discrete behavior itself, and the result of that behavior; critical incidents can be thought of as generally conforming to an "A–B–C" format (antecedent–behavior–consequence; Weekley, Ployhart, & Holtz, 2006). Critical incidents are usually elicited from SMEs for the job, domain, or task of interest; they should consist solely of specific circumstances, behaviors, and results that SMEs have personally witnessed, not hypothetical situations and actions, incidents that SMEs have merely "heard about," or general behavioral tendencies (e.g., comes to work early, is awkward when interacting with customers). As BARS aim to capture the entire behavioral continuum constituting performance, when critical incidents are being generated, not only should highly effective and ineffective behaviors be sampled but also "merely satisfactory" behaviors. Critical incidents can be gathered in many different ways, including in large workshops (Motowidlo et al., 1992), small groups (Hedge et al., 2004; Motowidlo, Dunnette, & Carter, 1990), individual interviews (Martin-Raugh, Kell, & Motowidlo, 2016), or through the review of archival materials (Weekley et al., 2006). At least several hundred critical incidents are usually gathered, with the goal being to reach the point where *saturation* occurs and new incidents feature largely redundant material. This *saturation point* is used as a heuristic to determine when the behavioral domain has been adequately covered.

Step 2

The BARS developers edit the critical incidents into a common format and remove redundancy before examining their content, identifying common themes, and developing job performance dimensions based on those themes. The developers create labels and definitions for these performance categories based on their content.

Step 3

Commonly referred to as *retranslation*, a second group of SMEs (not overlapping with the SMEs that generated the critical incidents) is given the critical incidents in randomized order, along with the list of performance categories and their definitions. The SMEs independently sort each critical incident into the performance category in which they believe it best fits.

Step 4

The BARS developers compute agreement statistics for each critical incident to determine which incidents SMEs agreed upon in their placements and which they did not. Incidents that do not meet some predetermined agreement standard (e.g., 80%) are discarded.

Step 5

A third group of SMEs (not overlapping with the first two groups) is provided with the critical incidents and the performance categories they belong to. These SMEs rate the incidents for effectiveness, often on a scale of 1 (*very ineffective*) to 7 (*very effective*).

Step 6

The BARS developers compute mean effectiveness values for each incident and examine the standard deviation of the SMEs' ratings to assess the degree to which the SMEs agreed upon the effectiveness of the incident. Incidents that do not meet some predetermined agreement standard (e.g., standard deviation of .5 or less) are discarded.

Step 7

Critical incidents that have survived both Step 4 and Step 6 are used to prepare the final BARS, with the mean effectiveness ratings of the incidents determining their placement on the continuum.

Modifications of the Basic Procedure

Deductive Versus Inductive

Step 2 as described constitutes an inductive approach to defining the performance domain, drawing solely on content commonalities of the critical incidents; SMEs' implicit ideas about what constitutes effective behavior are made explicit through their provision of critical incidents (Guion, 2011). However, a deductive approach can also be applied, wherein the domains of interest are first specified, even if vaguely, and critical incidents generated with the goal of concretizing and defining those domains. These initial dimensions might be derived from, for example, prior research and theory about job-general performance categories (e.g., Campbell, 1990, 2012) or SMEs' opinions about job-specific performance dimensions (e.g., Goodale & Burke, 1975; Zedeck, Imparato, Krausz, & Oleno, 1974).

Number of SMEs

When appropriate resources are available, additional SMEs may be recruited to examine the quality of the critical incidents initially gathered or to generate additional incidents to "fill out" inductively derived performance dimensions (Campbell et al., 1973). Alternatively, more SMEs can be used to extend the processes described previously, having multiple groups of SMEs independently retranslate the incidents and perhaps provide effectiveness ratings (Wollack, Goodale, Wijting, & Smith, 1971). SME groups may be purposefully chosen so that they differ in their perspective (e.g., organizational membership, rank within the same organization) on the job in question, to ascertain the extent to which effectiveness ratings and dimension assignments are generalizable (e.g., Borman & Vallon, 1974; Landy et al., 1976; Motowidlo & Borman, 1977; Motowidlo & Peterson, 2008; Zedeck et al., 1974).

Alternatively, when resources are scarce and fewer SME groups are available, the same SMEs may be called upon to perform multiple tasks; often this entails the same group of SMEs rating and retranslating critical incidents (Bernardin et al., 1976; Campbell et al., 1973; Hedge et al., 2004; Kavanagh & Duffy, 1978). The retranslation process may also be extremely truncated; in Campion, Pursell, and Brown (1988), an analyst independent of the project reclassified 9% of the critical incidents. Although not widely used, shortcut approaches have also been developed that, in essence, entail the same SMEs undertaking all of the major BARS construction steps: providing critical incidents, evaluating their effectiveness, and sorting them into performance dimensions (Champion et al., 1988; Green et al., 1981).

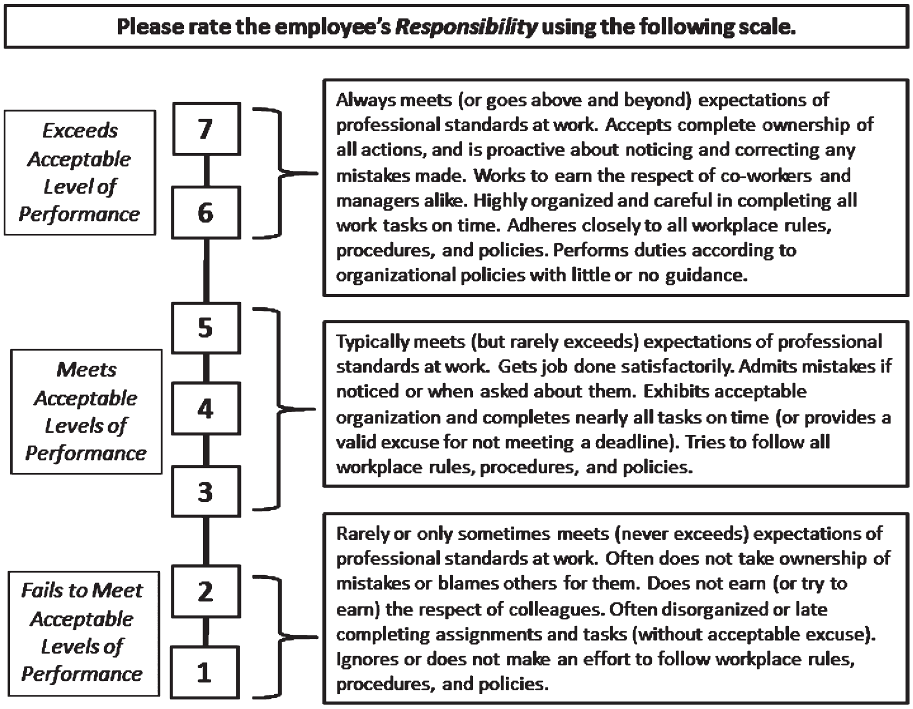
Additional Analyses

Critical incidents may be rated for more than effectiveness and have also been evaluated for frequency (Beatty, Schneier, & Beatty, 1977), importance (Dickinson & Tice, 1973), specificity (Kinicki & Bannister, 1988), and probability of occurrence (Zedeck, Kafry, & Jacobs, 1976). These ratings are sometimes subjected to factor analysis in addition to, or instead of, the judgmental content analysis described in Step 2 (Dickinson & Tice, 1973; Sprecher, 1965; Wollack et al., 1971). Rather than rating incidents for effectiveness, SMEs may be asked to provide paired comparisons of their effectiveness (Landy & Barnes, 1979). In order to assess the item discrimination of each incident, paired comparisons may be used to rate two groups of organizational incumbents, one identified as consisting of outstanding performers and the other identified as consisting of poor performers (Guion, 2011). Beyond individual incidents, entire performance dimensions can also be rated: for example, on their importance in determining overall job performance (Arvey & Hoyle, 1974).

Behavioral Summary Scales

Ironically, the concrete behavioral anchors that are often touted as a major advantage of BARS may also cause some raters difficulty due to their extreme specificity (Atkin & Conlon, 1978). The issue is stated succinctly by Borman (1979, p. 412): "The central problem is that raters often have difficulty discerning any behavioral similarity between a ratee's performance and the highly specific behavioral examples used to anchor the scale." Indeed, the same level of performance may be reached via different behavioral routes, but offering only a single example at each point on the rating scale may implicitly limit the rater's perspective on what constitutes effective performance at the level that point represents (Borman, Hough, & Dunnette, 1976). The behavioral summary scale (BSS; Borman, 1979, 1986) type of BARS was developed to address this issue. Creation of BSS entails taking critical incidents that have been reliably retranslated and reliably rated for effectiveness and further content analyzing them. Brief behavioral statements are then written that summarize the content that cuts across the critical incidents representing the same level of performance in each performance dimension. Although these statements are more general than individual critical incidents, they are still rooted in behaviors and are not nearly as abstract as the adjectives frequently used to assess personality traits.

Figure 2 consists of an example BSS. Examination of this figure reveals that BSSs differ not only in their use of behavioral statements (instead of critical incidents) to anchor scale points, but also that these behavioral statements anchor multiple scale points. The rationale for assigning statements to several scale points instead of single ones is that even when there is high agreement on the effectiveness level of an incident, there is likely enough variability in SMEs' judgments that reasonable raters may disagree on whether a behavior is "truly" indicative of, say, performance at an effectiveness level of *6* or an effectiveness level of *7* (Bernardin & Smith, 1981). To account for the possibility of these "flip-flops," summary statements are assigned as anchors to two or three scale points, which themselves are categorized assigned to a performance level (e.g., *exceeds expectations* versus *falls below expectations*). As raters may be able to make only coarse distinctions between ratees (Atkin & Conlon, 1978), this strategy also allows the possibility of later collapsing the numeric ratings so that final scores are assigned only to the three broad performance levels.


Please rate the employee's *Responsibility* using the following scale.
Exceeds Acceptable Level of Performance
7
6
Meets Acceptable Levels of Performance
5
4
3
Fails to Meet Acceptable Levels of Performance
2
1

Always meets (or goes above and beyond) expectations of professional standards at work. Accepts complete ownership of all actions, and is proactive about noticing and correcting any mistakes made. Works to earn the respect of co-workers and managers alike. Highly organized and careful in completing all work tasks on time. Adheres closely to all workplace rules, procedures, and policies. Performs duties according to organizational policies with little or no guidance.

Typically meets (but rarely exceeds) expectations of professional standards at work. Gets job done satisfactorily. Admits mistakes if noticed or when asked about them. Exhibits acceptable organization and completes nearly all tasks on time (or provides a valid excuse for not meeting a deadline). Tries to follow all workplace rules, procedures, and policies.

Rarely or only sometimes meets (never exceeds) expectations of professional standards at work. Often does not take ownership of mistakes or blames others for them. Does not earn (or try to earn) the respect of colleagues. Often disorganized or late completing assignments and tasks (without acceptable excuse). Ignores or does not make an effort to follow workplace rules, procedures, and policies.

Figure 2 Example behavioral summary scale (after Borman, Hough, & Dunnette, 1976).

Behaviorally Anchored Rating Scales for Evaluating Interview Performance

The preceding makes it clear that the procedures for developing BARS for job performance appraisal are well established. Surprisingly, much less attention has been paid to techniques for creating BARS for evaluating interview performance. In contrast to the copious documentation often provided in articles describing the creation of BARS for job performance appraisal, the documentation is minimal for BARS for interview performance appraisal. The following quotes are representative: "The job experts, together with the first author, reviewed the questions and attempted to generate anchors for the three scale points. Discussion ensued concerning each suggested anchor until the four experts came to agreement regarding the best anchor for the scale point in question" (Weekley & Gier, 1987, p. 485) and "For each of the 25 interview questions … benchmark answers (1 = *poor*, 3 = *average*, 5 = *good*) were developed by the supervisors of the office clerical employees" (Latham & Saari, 1984, p. 571).

There is no standard method for generating the content used to create interview BARS (Campion et al., 1997). Some of the methods that have been used include using answers heard by SMEs during job interviews from candidates who were actually hired (Latham et al., 1980), having interviewers develop anchors from notes about responses to similar questions (Green, Alter, & Carr, 1993), brainstorming and speculation by SMEs (Campion et al., 1988; Robertson, Gratton, & Rout, 1990), and having project staff draft a first version of the BARS and revising it after consulting with SMEs (M. A. Campion, personal communication, January 27, 2016). BARS can be generated for questions tied to a critical incident analysis of a specific job (Motowidlo et al., 1992) or for preexisting interview questions (Latham & Saari, 1984). Because the anchors are tied closely to specific questions, they are usually not subjected to the retranslation process described in Step 3 (Green et al., 1993). A recent review of the employment interview literature well sums up the situation regarding BARS for evaluating interview performance: "Despite the importance of [B]ARSs to structured interviews, the science of rating scales is still surprisingly underresearched" (Levashina et al., 2014, p. 274).

The Current Study

The current investigation explores the possibility of increasing the efficiency of BARS construction by gathering responses to structured interview questions through an online crowdsourcing program, AMT. This approach is particularly amenable to developing materials to evaluate interview performance: The constructs assessed in many interviews are highly general and relevant to many jobs, suggesting that the traditional notion that there are SMEs for them may not fully apply. Indeed, novices have been shown to be capable of earning scores similar to those of experienced managers on a job knowledge test tapping interpersonal constructs (e.g., communication, leadership; Motowidlo & Beier, 2010). Adopting a broader definition of "expert" means a wider population can viably provide the basic materials used to construct BARS for rating interview performance—and online crowdsourcing platforms are an efficient means for accessing this population.

Additionally, it is an opportune time to explore strategies to ease the construction of structured interviews. Although interviews have always been popular, their popularity may only increase in the future due to the rising focus on noncognitive skills (Kyllonen, 2013; Kyllonen, Lipnevich, Burrus, & Roberts, 2014; Naemi, Burrus, Kyllonen, & Roberts, 2012). Interest in noncognitive skills has greatly expanded over the past two decades, despite significant concerns about the adequacy of their measurement (Duckworth & Yeager, 2015). Structured interviews are viable measures of noncognitive skills (Huffcutt et al., 2001) and tend to exhibit respectable psychometric characteristics, suggesting they could be used on a wide scale to assess these constructs.

We first describe the development of 12 interview questions to capture the four applied social skills reviewed. Next, the gathering and preparation of data from AMT respondents are detailed. Finally, we describe the creation of BARS from these AMT responses. We reasoned that interviewees can provide different kinds of answers that are of comparable quality, just as people can reach the same level of performance in different ways, and thus chose to create BSS instead of traditional BARS.

As far as we are aware, this is the first time crowdsourced text will be used to construct a performance appraisal tool of any kind.

Method

Part 1: Procedure for the Development of Interview Questions

The second and third authors reviewed the job interview research literature and other sources to locate structured interview questions they judged to be indicative of four applied social skills identified (Huffcutt et al., 2001): communication, leadership, negotiation and persuasion, and teamwork. Based on these sources, they developed three PBQs targeting each construct to enhance the reliability of each construct's assessment. Each question ends with the instruction "Please provide details about the background of the situation, the behaviors you carried out in response to that situation, and what the outcome was," the goal being to encourage participants to provide their responses in the form of critical incidents. The 12 interview questions are available in Appendix A.

Part 2: Behavioral Summary Scale Development

Participants

The sample comprised 68 individuals recruited via AMT. Subjects were paid $3 for contributing to the study, which took approximately 30 minutes to complete. Participants' average age was 34 years ($SD = 10.49$). The sample was 50% female, 48.5% male, and one individual (1.5%) who did not specify a sex. In terms of race and ethnicity, the composition of the sample was 42.6% White (non-Hispanic); 35.3% "Other American;" 7.4% Black or African American; 5.9% Asian or Asian American; 2.9% Puerto Rican; and 1.5% Mexican, Mexican American, or Chicano. One individual (1.5%) preferred not to provide ethnicity information. Sixty-six individuals (97.1%) indicated that they felt most comfortable communicating in English. Participants were not screened according to their educational level, which consisted of the following: grade school or less (2.9%), some high school (13.2%), business or trade school (4.4%), some college (26.5%), associate's or 2-year degree (10.3%), bachelor's or 4-year degree (33.8%), some graduate or professional degree (2.9%), and graduate or professional degree (5.9%). Participants had an average of 3.51 years of work experience ($SD = 1.65$). The Occupational Information Network (O*NET; Peterson, Mumford, Borman, Jeanneret, & Fleishman, 1999) job families represented by respondents' occupations were 10.3% Computer & Mathematical; 10.3% Sales and Related; 14.7% Office and Administrative Support; 16.2% Management; 7.4% Arts, Design, Entertainment, Sports, and Media; 4.4% Business and Financial Operations; 1.5% Healthcare Practitioners and Technical; 1.5% Healthcare Support; 4.4% Education, Training, and Library; 2.9% Architecture and Engineering; 1.5% Construction and Extraction; 1.5% Protective Service; 1.5% Personal Care and Service; and

21.9% Other. Respondents' average cumulative grade point average was 2.57 on a 4-point scale (*SD* = 0.87). For the 21 individuals (30.9%) who reported a composite SAT score, the mean was 1245.48 (*SD* = 163.75). Of the 15 individuals (22.1%) who reported a composite ACT score, the mean was 26.47 (*SD* = 4.45).

Procedure

After consenting to participate in the study, subjects were asked a series of background questions. Then they were told, "You will be presented with a series of questions that will ask you to recall a specific event or accomplishment from your past. Please respond to these questions honestly, and provide as much detail as you can by typing in the text boxes displayed below each question." Next, participants were presented with the 12 PBQs and typed their responses into a text box provided for each question.

AMT participants provided 681 responses—84% of the total number of responses (816) that could have been provided; 58 of the 68 respondents (85%) answered all 12 interview questions. Preliminary review removed unusable responses (e.g., irrelevant, nonsensical) and reduced this pool to 652 usable responses, present in forms largely conforming to the structure of critical incidents. These critical incidents were edited for concision, grammar, and redundancy. In addition, the outcome of the focal behavior was removed so that we could collect judgments of incidents' effectiveness based on the behavior described in the incident, rather than on its results (cf. Kell, Motowidlo, Martin, Stotts, & Moreno, 2014; Motowidlo, Crook, Kell, & Naemi, 2009), consonant with behavior-based theories of job performance (Motowidlo & Kell, 2013). The first three authors rated the effectiveness of each incident on a 7-point scale, where 1 = *very ineffective* and 7 = *very effective*. The intraclass correlation coefficient (ICC) for the three raters across the 12 questions was .74. Per dimension, the ICCs were as follows: Communication (.75), Leadership (.66), Persuasion and Negotiation (.82), and Teamwork (.72).

Results

We used the standard deviation of the effectiveness ratings for each incident to index the extent of agreement among the SMEs, as is traditionally done when developing BARS (Schwab et al., 1975). Incidents with high agreement (a standard deviation of 1.5 or less) were retained and the others discarded. Incidents receiving a mean expert rating of 3 or less were considered indicative of the low range of effectiveness. Incidents with a mean effectiveness rating over 3 but under 5 were considered indicative of the medium effectiveness range, and incidents with mean effectiveness ratings of 5 or more were considered indicators of high levels of effectiveness. (It is worth noting that benchmarks for agreement and effectiveness should be adjusted accordingly if the number of points in the effectiveness scale is compressed or expanded in future iterations.) Two examples of critical incidents for each effectiveness range (with the construct each represents) follow:

High Effectiveness

> I was working with coworkers on a project translating the employee handbook to Spanish. There wasn't a formal leader, but I decided to start by doing a significant amount of work, then giving what I had done to the other members for any editing or correction. (Leadership)
>
> We had a meeting to discuss how some of the database tables work and how reports are produced from them. I had to explain to about ten people how this works. I tried my best to use nontechnical language and encouraged them to ask questions; this way they could learn in their own words instead of my technical jargon. (Communication)

Medium Effectiveness

> I am often in a situation where I work with someone I dislike or do not get along with, but each time I have to, I remind myself that this is what I chose as my career and that I must act as professionally as I can toward that person and the company that person is working for. I don't take things personally and I try to keep a neutral stance and mood toward the person I am called to deal with. (Teamwork)

> Our union contract was almost up. Everyone wanted a raise, but things were not looking good for the company. I was the unit steward, and I had to tell everyone the facts. I proposed we agree to the wage freeze and keep our health insurance benefits. (Persuasion/Negotiation)

Low Effectiveness

> A coworker of mine I did not necessarily get along with. I felt like she was always undermining me and not listening when I have the expertise in the department. I tried to just not talk to her other than when we were forced to have interactions. (Teamwork)
>
> When I was working my silly retail job, there were always people who didn't get along. Two in particular, a girl in her early 20s and a guy nearing 40, come to mind. Both of them were terrible people in my mind, so I made no effort to mediate. (Leadership)

The first four authors independently examined the remaining critical incidents and identified common themes they believed defined performance for each effectiveness range for each of the 12 questions (Borman, 1979; Hedge et al., 2004). Over the course of several meetings, they worked together to write and reach consensus on summary statements that captured these common themes. They wrote a total of 108 summary statements, allotting three each to anchor each performance range for each question. Anchors between the effectiveness extremes depict a progression of increasingly effective performance. The final BSS for each of the 12 questions comprise Appendix B.

Discussion

Appropriately structuring job interviews is essential for enhancing the reliability, validity, and fairness of their scores (Levashina et al., 2014). One important element of interview structure is the provision of standardized, anchored rating scales for the evaluation of interviewees' responses. This report examined the feasibility of constructing a type of BARS (BSS) for the evaluation of interview performance using responses gathered through an online crowdsourcing platform, AMT. These participants were not "experts" in the traditional sense of having significant experience with a particular job or occupation, as the constructs assessed by the interview questions were highly general.

Of the total 681 critical incidents provided, 652 (95.7%) were of sufficient quality (e.g., comprehensible in terms of the behavior depicted, relevant to the question they were answering) to be initially considered for serving as the basis for behavioral summary statements. This result is encouraging considering concerns about the quality of the data produced by crowdsourced respondents (Hamby & Taylor, 2016) and the fact that, because this was an exploratory investigation, we did not screen participants on any variables (e.g., education, work experience). A single study of a single sample does not constitute definitive evidence for the conclusion that critical incidents of usable quality can consistently be generated by crowdsourced workers. Nonetheless, these results do suggest that further exploration of this option is worth pursuing, as it has significant practical advantages over traditional methods used to gather the materials needed to construct BARS, which have been acknowledged for decades as time-consuming and expensive to develop (Landy & Farr, 1980; Landy et al., 1976; Maurer, 1997).

Gathering the critical incidents that form the basis for BARS via crowdsourcing reduces the burden of only a single step in the BARS development process, but the amount of resources that can be expended in that step should not be underestimated. Critical incidents are often gathered personally by project staff at workshops that can take months to organize and a day or more to conduct (Dunnette, 1976). SMEs' and project staffers' schedules must be aligned and a location large enough to hold the workshop found. Money may have to be spent renting that space, providing meals and refreshments during the workshop, and reimbursing staff members for travel and parking costs. SMEs often have to be paid a substantial amount for them to be willing to give up vacation or weekend days to participate in the workshop. If the organization gathering the incidents does not have a sufficient number of laptops, SMEs will have to write out their incidents by hand, which staff members will then have to transcribe.

Much of this time and expense can be eliminated by gathering critical incidents through online crowdsourcing. Indeed, even if future investigations determine that crowdsourced participants do not consistently provide critical incidents of sufficient quality, or that BARS developed from their responses lack reliability, validity, or both, this study simply demonstrates the practical benefits of gathering critical incidents electronically. Indeed, if unscreened, anonymous, minimally compensated AMT workers are capable of providing a large number of viable critical incidents, presumably the results will be even better if the same method is used with samples of SMEs who have personal contact with project staff and are appropriately compensated.

Limitations and Recommendations for Future Research

We acknowledge the limitations of this investigation in the following sections. We integrate the discussion of this study's limitations with proposals for future research on the feasibility of developing BARS using materials gathered with online crowdsourcing platforms.

Background Characteristics

Crowdsourcing platforms offer many options for screening and selecting participants according to users' needs. Future studies should explore the impact of tailoring the nature of the sample providing critical incidents so it is well aligned with the incumbents of the job for which hiring is being conducted. As this was an exploratory effort, we did not screen AMT participants on any variables, including educational attainment and work-relevant variables such as job experience and organizational tenure. Future research should examine what impact, if any, various background variables have on the relevance and quality of the data produced. For example, do respondents with more work experience generate critical incidents that are more reliably rated for effectiveness and retranslated than respondents with less experience? In this study, the ICC for Leadership was rather low (.66); perhaps focusing only on online participants with a substantial amount of job experience would have yielded more unequivocally effective or ineffective critical incidents for this dimension.

For these reasons, we recommend creating a more comprehensive background questionnaire than the one used in this investigation. Such a questionnaire might include some or all of the following items: current job title, number of years in current job, number of years in industry, number of years in current organization, number of years of overall work experience, high school GPA, undergraduate GPA, major of highest degree (if attended college), institution attended for highest degree, and general type of organization working in (e.g., for profit, nonprofit, government, other). Additionally, the fact that crowdsourcing platforms rely on self-report is problematic because responses to background items can easily be faked, although this is less of a concern due to the low-stakes setting. Nonetheless, it may be worthwhile to ask participants to provide a link to a resource that allows for external verification of some of the background information they provide, such as a LinkedIn profile, company website, or personal website with a CV or resume. Many participants will surely decline to provide such information, but a benefit of online crowdsourcing platforms is the sheer number of individuals that participate in them: AMT alone has approximately 500,000 users (AMT, 2016). This means that a large number of incidents could still be obtained, even if they are gathered only from respondents willing to provide verifiable evidence of their qualifications. A downside of this technique, of course, is the possibility of sampling bias influencing the content of the incidents provided by this highly select group.

Amazon Mechanical Turk Behaviorally Anchored Ratings Versus Traditionally Developed Behaviorally Anchored Rating Scales

A single sample of AMT participants providing viable responses to behavioral interview questions in the form of critical incidents does not constitute sufficiency for using this approach to construct BARS for the purpose of evaluating interview or job performance. Additional studies must be conducted to demonstrate that AMT (or other crowdsourced) respondents consistently provide responses of sufficient quality to build BARS. Once this research is conducted, it will then be essential to examine the psychometric quality of BARS constructed from crowdsourced data.

As noted previously, what it "means" to be an SME is ambiguous for the broad, job-general constructs tapped by the structured interview used in this study. Nonetheless, it stands to reason that individuals with more overall job experience (note that our participants had fewer than 4 years of experience, on average) are better judges of what constitutes effective and ineffective behavior in "the workplace in general," even for such a broad domain as applied social skills. Indeed, although novices' judgments about the effectiveness of interpersonal behaviors are predictive of actual job performance, experienced managers' judgments are *more* predictive (Motowidlo & Beier, 2010). For example, relatively inexperienced respondents may be unaware of certain critical behaviors that are associated with important workplace outcomes, due to a lack of work experience. Consequently, these behaviors would be excluded from the content of the BARS, which could degrade their validity for predicting job performance. If inexperienced respondents are unaware of, or unable to focus sufficient attention on, certain work behaviors that play a key role in determining workplace outcomes, this may lead them to provide a larger number of incidents that focus on less important behaviors that are, consequently, more equivocal in their effectiveness. If incorporated into BARS, these ambiguous behaviors could produce more disagreement among raters, decreasing the reliability of the instrument and, as a byproduct, negatively influencing predictive validity.

Future Directions

We have made clear that future investigations into the properties of BARS generated through crowdsourced data must be conducted before such respondents can be seriously considered as sources of data for developing this type of measure. Now, we broadly outline five research designs that can be used to guide these future investigations. These designs are intended to more rigorously assess both the quality of crowdsourced interview data and to evaluate it in comparison with the quality of the data provided by SMEs (e.g., experienced interviewers or job incumbents). In this investigation, “quality” was subjectively assessed by project staff in terms of the intelligibility, coherence, and relevance of the data provided by AMT workers. In the future, it will be important to examine quality not only in terms of the content of the data, but also the psychometric characteristics of the scales that result from them.

Design 1

A group of SMEs reviews the critical incidents generated for each question and rates how *valid*, *relevant*, or *acceptable* they are as answers to the questions they were produced in response to. Only critical incidents with mean scores that pass some predetermined threshold of “acceptability” (e.g., above the midpoint of the rating scale used) and agreement among the judges (e.g., 70%) are considered viable resources for constructing BARS.

Design 2

Ask SMEs to rate crowdsourced incidents for effectiveness. Also provide SMEs with all of the interview questions and ask them to allocate each critical incident to the interview question they believe it most realistically could have been provided as an answer to (i.e., retranslate the answers). Develop BARS only from critical incidents reliably rated for effectiveness *and* retranslated according to predetermined standards.

Design 3

Gather critical incidents from a crowdsourced sample and an SME sample. Present these incidents to a second, nonoverlapping group of SMEs in random order and ask them to guess which incidents were provided by SMEs versus online respondents. Examine whether SMEs are, overall, able to guess who generated the incidents at above chance.

Design 4

Combine and expand on Designs 1 and 3. Take incidents provided by online and SME samples, provide them to a second group of SMEs, and ask this group to rate them for *effectiveness*, *quality*, or *favorability* (e.g., “How comfortable would you be using this incident to evaluate the performance of actual people in job interviews?”) and retranslate them. Examine whether any of these evaluations differ statistically and practically according to the source of the incidents.

Design 5

Develop separate BARS from totally independent sets of incidents created by separate groups of online participants and SMEs. Have two (or more) raters use both sets of BARS to evaluate the performance of videotaped real or mock interviewees and examine whether the psychometric characteristics of the scores (e.g., interrater reliability, intercorrelations among the dimensions) differ. If possible, have the raters rate the interview videos again at least one month later to assess differences in test–retest reliability and interrater reliability across time. If job performance ratings or other criteria (e.g., absences, awards, disciplinary actions) are available, examine whether there are statistical and practical differences between the validity coefficients for predicting these criteria. If adequate resources are available, also examine the convergent and discriminant validity of the two sets of BARS with correlates of interview performance (e.g., cognitive ability scores, personality ratings, SJTs).

Even if BARS generated through crowdsourcing are eventually shown to be comparable to SME-generated BARS for evaluating interview performance, this does not imply that crowdsourced BARS are a viable means of evaluating job performance. Job performance appraisal BARS are often created from the critical incidents obtained through a systematic job analysis using SMEs from a single job, and it is difficult to see how this task could be appropriately "outsourced" to online respondents. Indeed, the source of BARS being job experts' judgments automatically imbues them with job relevance, which is a critical component of legal defensibility in adverse impact cases (Jacobs et al., 1980; Jeanneret & Zedeck, 2010; Landy, Gutman, & Outtz, 2010). Jacobs et al. (1980) noted:

> Perhaps the strongest attribute of the BARS methodology is its ability to yield job analysis information performed by the people who know the job best and written in their language. By generating and evaluating behavioral items necessary for the final format, the BARS methodology results in explicit statements regarding requisite job behaviors and their perceived value. On this level BARS item generation can be seen as meeting the criterion of relevancy. (p. 606)

The fact that experts are heavily involved in the creation of BARS for job performance appraisal is one of the major reasons they continue to be a popular means of performance appraisal, despite equivocal evidence for them being superior to other methods of job performance appraisal that are similarly rigorously developed (Borman, 1991).

Conclusion

This study was an exploratory investigation of the viability of using responses to structured interview questions gathered through a crowdsourcing platform to create BARS. Respondents were not screened on education- or work-related variables. Nonetheless, the results of this study suggest that online respondents are willing to produce, and capable of producing, responses of sufficient quality to create BARS for evaluating applied social skills performance in an interview. Future research should strive to replicate and extend these results by comparing the psychometric properties of BARS generated from crowdsourced materials to SME-created BARS.

Note

[1] The original scale now generally referred to as BARS was called a behavioral expectations scale (BES; Smith & Kendall, 1963). As variations in methodology and format arose over time, the generic term *BARS* came to refer to a family of measures that includes BES, behavioral observation scales, behavioral discrimination scales, and behavioral summary scales.

References

Amazon Mechanical Turk. (2016). *Service summary*. Retrieved from https://requester.mturk.com/tour

Anastasi, A. (1938). Faculties versus factors: A reply to Professor Thurstone. *Psychological Bulletin*, *35*, 391–395.

Arvey, R. D., & Hoyle, J. C. (1974). A Guttman approach to the development of behaviorally based rating scales for systems analysts and programmer/analysts. *Journal of Applied Psychology*, *59*, 61–68.

Atkin, R. S., & Conlon, E. J. (1978). Behaviorally anchored rating scales: Some theoretical issues. *Academy of Management Review*, *3*, 119–128.

Bass, B. M. (1992). *Bass and Stogdill's handbook of leadership* (3rd ed.). New York, NY: The Free Press.

Beatty, R. W., Schneier, C. E., & Beatty, J. R. (1977). An empirical investigation of perceptions of ratee behavior frequency and ratee behavior change using behavioral expectation scales (BES). *Personnel Psychology*, *30*, 647–658.

Bernardin, H. J., LaShells, M. B., Smith, P. C., & Alvares, K. M. (1976). Behavioral expectation scales: Effects of developmental procedures and formats. *Journal of Applied Psychology*, *61*, 75–79.

Bernardin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored ratings scales (BARS). *Journal of Applied Psychology*, *66*, 458–463.

Blood, M. R. (1974). Spin-offs from behavioral expectation scale procedures. *Journal of Applied Psychology*, *59*, 513–515.

Bolanvich, D. J. (1994). Selection of female engineering trainees. *Journal of Educational Psychology*, *35*, 545–553.

Bono, J. E., & Judge, T. A. (2004). Personality and transformational and transactional leadership: A meta-analysis. *Journal of Applied Psychology*, *89*, 901–910.

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

Borman, W. C. (1986). Behavior-based rating scales. In R. A. Beck (Ed.), *Performance assessment: Methods and applications* (pp. 100–120). Baltimore, MD: Johns Hopkins University Press.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial & organizational psychology* (Vol. *2*, pp. 271–326). Palo Alto, CA: Consulting Psychologists Press.

Borman, W. C., & Dunnette, M. D. (1975). Behavior-based versus trait-oriented performance ratings: An empirical study. *Journal of Applied Psychology*, *60*, 561–565.

Borman, W. C., Hough, L. M., & Dunnette, M. D. (1976). *Development of behaviorally based rating scales for evaluating the performance of U.S. Navy recruiters* (NPRDC Technical Report 76–31). San Diego, CA: Navy Personnel Research and Development Center.

Borman, W. C., & Vallon, W. R. (1974). A view of what can happen when behavioral expectation scales are developed in one setting and used in another. *Journal of Applied Psychology*, *59*, 197–201.

Brungardt, C. (2011). The intersection between soft skill development and leadership education. *Journal of Leadership Education*, *10*, 1–22.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5.

Campbell, J. P. (1990) Modeling the performance prediction problem in industrial and organizational psychology. In M. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. *1*, pp. 687–731). Palo Alto, CA: Consulting Psychologists Press.

Campbell, J. P. (2012). Behavior, performance, and effectiveness in the 21st century. In S. W. J. Kozlowski (Ed.), *The Oxford handbook of organizational psychology* (pp. 159–195). New York, NY: Oxford University Press.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. (1973). The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology*, *57*, 15–22.

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

Campion, M. A., Pursell, E. D., & Brown, B. K. (1988). Structured interviewing: Raising the psychometric properties of the employment interview. *Personnel Psychology*, *41*, 25–42.

Cantor, N. (1990). From thought to behavior: "Having" and "doing" in the study of personality and cognition. *American Psychologist*, *45*, 735–750.

Carnevale, A. P., Gainer, L. J., & Meltzer, A. S. (1990). *Workplace basics: The essential skills employers want*. San Francisco, CA: Jossey-Bass.

Carnevale, A. P., Smith, N., & Strohl, J. (2013). *Recovery: Job growth and education requirements through 2020*. Washington, DC: Georgetown Public Policy Institute, Center on Education and the Workforce.

Casner-Lotto, J., & Barrington, L. (2006). *Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. workforce*. New York, NY: The Conference Board, Partnership for 21st Century Skills, Corporate Voices for Working Families, & Society for Human Resources Management.

Champion, C. H., Green, S. B., & Sauser, W. I. (1988). Development and evaluation of shortcut- derived behaviorally anchored rating scales. *Educational and Psychological Measurement*, *48*, 29–41.

Dalessio, A. T., & Silverhart, T. A. (1994). Combining biodata test and interview information: Predicting decisions and performance criteria. *Personnel Psychology*, *47*, 303–315.

DeGroot, T., & Kluemper, D. (2007). Evidence of predictive and incremental validity of personality factors, vocal attractiveness and the situational interview. *International Journal of Selection and Assessment*, *15*, 30–39.

DeGroot, T., & Motowidlo, S. J. (1999). Why visual and vocal interview cues can affect interviewers' judgments and predict job performance. *Journal of Applied Psychology, 84*, 986–993.

Dickinson, T. L., & Tice, T. E. (1973). A multitrait-multi method analysis of scales developed by retranslation. *Organizational Behavior and Human Performance*, *9*, 421–438.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*, 237–251.

Dunnette, M. D. (1976). Aptitudes, abilities, and skills. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 473–520). Chicago, IL: Rand McNally.

Durlach, P. J., Wansbury, T. G., & Wilkinson, J. G. (2008). *Cultural awareness and negotiation skills training: Evaluation of a prototype semi-immersive system*. Orlando, FL: U.S. Army Research Institute for the Behavioral and Social Sciences.

Eley, M. G., & Stecher, E. J. (1997). A comparison of two response scale formats used in teaching evaluation questionnaires. *Assessment & Evaluation in Higher Education*, *22*, 65 – 79.

End, C. M., & Saunders, K. (2013). Powerless and jobless? Comparing the effects of powerless speech and speech disorders on an applicant's employability. *Frontiers*, *2*, 4 – 9.

Finegold, D., & Notabartolo, A. S. (2010). *21st century competencies and their impact: An interdisciplinary literature review.* Report commissioned for the NRC Project on Research on 21st Century Competencies. Retrieved from http://www.hewlett.org/uploads/21st_Century_Competencies_Impact.pdf

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

Gardner, H. (1993). *Frames of mind*. New York, NY: Basic Books.

Gerstner, C. R., & Day, D. V. (1997). Meta-analytic review of leader – member exchange theory: Correlates and construct issues. *Journal of Applied Psychology*, *82*, 827 – 844.

Gibbons, P., Busch, J., & Bradac, J. J. (1991). Powerful versus powerless language: Consequences for persuasion, impression formation, and cognitive response. *Journal of Language and Social Psychology, 10*, 115 – 133.

Goodale, J. G., & Burke, R. J. (1975). Behaviorally based rating scales need not be job specific. *Journal of Applied Psychology*, *60*, 389 – 391.

Green, P. C., Alter, P., & Carr, A. F. (1993). Development of standard anchors for scoring generic past-behaviour questions in structured interviews. *International Journal of Selection and Assessment*, *1*, 203 – 212.

Green, S. B., Sauser, W. I., Fagg, J. N., & Champion, L. C. H. (1981). Shortcut methods for deriving behaviorally anchored rating scales. *Educational and Psychological Measurement*, *41*, 761 – 775.

Guion, R. M. (2011). *Assessment, measurement, and prediction for personnel decisions* (2nd ed.). New York, NY: Routledge.

Hamby, T., & Taylor, W. (2016). Survey satisficing inflates reliability and validity measures: An experimental comparison of college and Amazon Mechanical Turk samples. *Educational and Psychological Measurement, 76*(6), 912 – 932. . https://doi.org/10.1177/0013164415627349

Hartsough, C. S., Perez, K. D., & Swain, C. L. (1998). Development and scaling of a preservice teacher rating instrument. *Journal of Teacher Education*, *49*, 132 – 139.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57,* 639 – 683.

Hedge, J. W., Borman, W. C., Bruskiewicz, K. T., & Bourne, M. J. (2004). The development of an integrated performance category system for supervisory jobs in the U.S. Navy. *Military Psychology*, *16*, 231 – 243.

Hoffman, C. C., & Holden, L. M. (1993). Dissecting the interview: An application of generalizability analysis. In D. L. Denning (Chair), *Psychometric analysis of the structured interview.* Symposium conducted at the 8th Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Hollandsworth, J. G., Kazelskis, R., Stevens, J., & Dressel, M. E. (1979). Relative contributions of verbal, articulative, and nonverbal communication to employment decisions in the job interview setting. *Personnel Psychology, 32*, 359 – 367.

Hom, P. W., DeNisi, A. S., Kinicki, A. J., & Bannister, B. D. (1982). Effectiveness of performance feedback from behaviorally anchored rating scales. *Journal of Applied Psychology*, *67*, 568 – 576.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, *86*, 897 – 913.

Huffcutt, A. I., & Roth, P. L. (1998). Racial group differences in employment evaluations. *Journal of Applied Psychology*, *83*, 179 – 189.

Jacobs, R., Kafry, D., & Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. *Personnel Psychology*, *33*, 595 – 640.

Jain, S., & Anjuman, A. S. S. (2013). Facilitating the acquisition of soft skills through training. *IUP Journal of Soft Skills*, *7*, 32 – 39.

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

Jeanneret, P. R., & Zedeck, S. (2010). Professional guidelines/standards. In J. Farr & N. Tippins (Eds.), *Handbook of employee selection* (pp. 593 – 625). New York, NY: Taylor & Francis Group.

Jordan, P. J., Ashkanasy, N. M., Härtel, C. E., & Hooper, G. S. (2002). Workgroup emotional intelligence: Scale development and relationship to team process effectiveness and goal focus. *Human Resource Management Review*, *12*, 195 – 214.

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology*, *87*, 765 – 780.

Kavanagh, M. J. (1971). The content issue in performance appraisal: A review. *Personnel Psychology*, *24*, 653 – 668.

Kavanagh, M. J., & Duffy, J. F. (1978). An extension and field test of the retranslation method for developing rating scales. *Personnel Psychology*, *31*, 461 – 470.

Kell, H. J., Motowidlo, S. J., Martin, M. P., Stotts, A. L., & Moreno, C. A. (2014). Testing for independent effects of prosocial knowledge and technical knowledge on skill and performance. *Human Performance*, *27*, 311 – 327.

Klehe, U. C., & Latham, G. P. (2005). The predictive and incremental validity of the situational and patterned behavior description interviews for teamplaying behavior. *International Journal of Selection and Assessment*, *13*, 108 – 115.

Kinicki, A. J., & Bannister, B. D. (1988). A test of the measurement assumptions underlying behaviorally anchored rating scales. *Educational and Psychological Measurement*, *48*, 17 – 27.

Kyllonen, P. C. (2013). Soft skills for the workplace. *Change*, *45*, 16 – 23.

Kyllonen, P. C., Lipnevich, A. A., Burrus, J., & Roberts, R. D. (2014). *Personality, motivation, and college readiness: A prospectus for assessment and development* (Research Report No. RR-14-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12004

Landy, F. J. (1976). The validity of the interview in police officer selection. *Journal of Applied Psychology, 61,* 193 – 198.

Landy, F. J., & Barnes, J. L. (1979). Scaling behavioral anchors. *Applied Psychological Measurement*, *3*, 193 – 200.

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

Landy, F. J., Farr, J. L., Saal, F. E., & Freytag, W. R. (1976). Behaviorally anchored scales for rating the performance of police officers. *Journal of Applied Psychology*, *61*, 750 – 758.

Landy, F. J., & Guion, R. M. (1970). Development of scales for the measurement of work motivation. *Organizational Behavior and Human Performance*, *5*, 95 – 103.

Landy, F. J., Gutman, A., & Outtz, J. (2010). A sampler of legal principles in employment selection. In J. Farr & N. Tippins (Eds.), *Handbook of employee selection* (pp. 627 – 676). New York, NY: Taylor & Francis Group.

Latham, G. P., & Saari, L. M. (1984). Do people do what they say? Further studies on the situational interview. *Journal of Applied Psychology*, *69*, 569 – 573.

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

Lee, M. A., & Mather, M. (2008). U.S. labor force trends. *Population Bulletin, 63*, 1 – 20.

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology*, *67*, 241 – 293.

Martin-Raugh, M. P., Kell, H. J., & Motowidlo, S. J. (2016). Prosocial knowledge mediates effects of agreeableness and emotional intelligence on prosocial behavior. *Personality and Individual Differences*, *90*, 41 – 49.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*, 1 – 23.

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

Mayfield, E. C. (1964). The selection interview: A re-evaluation of published research. *Personnel Psychology*, *17*, 239 – 260.

McCall, M. W., & Bobko, P. (1990). Research methods in the service of discovery. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. *1*, pp. 381 – 418). Palo Alto, CA: Consulting Psychologists Press.

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.

Medical University of South Carolina. (2015). *University HR*. Retrieved from http://academicdepartments.musc.edu/hr/university

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13 – 103). New York, NY: Macmillan.

Morgeson, F. P., Reider, M. H., & Campion, M. A. (2005). Selecting individuals in team settings: The importance of social skills, personality characteristics, and teamwork knowledge. *Personnel Psychology*, *58*, 583 – 611.

Motowidlo, S. J., & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology*, *95*, 321 – 333.

Motowidlo, S. J., & Borman, W. C. (1977). Behaviorally anchored scales for measuring morale in military units. *Journal of Applied Psychology*, *62*, 177 – 183.

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

Motowidlo, S. J., Crook, A. E., Kell, H. J., & Naemi, B. (2009). Measuring procedural knowledge more simply with a single-response situational judgment test. *Journal of Business and Psychology*, *24*, 281 – 288.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, *75*, 640 – 647.

Motowidlo, S. J., & Kell, H. J. (2013). Job performance. In N. W. Schmitt & S. Highhouse (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., pp. 82 – 103). Hoboken, NJ: Wiley.

Motowidlo, S. J., & Peterson, N. G. (2008). Effects of organizational perspective on implicit trait policies about correctional officers' job performance. *Human Performance*, *21*, 396 – 413.

Muck, P. M., Hell, B., & Höft, S. (2008). Application of the principles of behaviorally anchored rating scales to assess the Big Five personality constructs at work. In J. Deller (Ed.), *Research contributions to personality at work*. Munich, Germany: Rainer Hampp Verlag.

Naemi, B., Burrus, J., Kyllonen, P. C., & Roberts, R. D. (2012). *Building a case to develop noncognitive assessment products and services targeting workforce readiness at ETS* (Research Memorandum No RM-12-23). Princeton, NJ: Educational Testing Service. http://www.ets.org/s/workforce_readiness/pdf/rm_12_23.pdf

National Research Council. (2011). *Assessing 21st century skills: Summary of a workshop*. Washington, DC: National Academies Press.

Ohland, M. W., Loughry, M. L., Woehr, D. J., Bullard, L. G., Felder, R. M., Finelli, C. J., … Schmucker, D. G. (2012). The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self- and peer evaluation. *Academy of Management Learning & Education*, *11*, 609–630.

Page, D., & Mukherjee, A. (2009). Effective technique for consistent evaluation of negotiation skills. *Education*, *129*, 521–533.

Parton, S. R., Siltanen, S. A., Hosman, L. A., & Langenderfer, J. (2002). Employment interview outcomes and speech style effects. *Journal of Language and Social Psychology, 21*, 144–161.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., & Fleishman, E. A. (Eds.). (1999). *An occupational information system for the 21st century: The development of O*NET*. Washington, DC: American Psychological Association.

Phillips, J. F. (1993). Predicting negotiation skills. *Journal of Business and Psychology, 7,* 403–411.

Reilly, N. P., Bocketti, S. P., Maser, S. A., & Wennet, C. L. (2006). Benchmarks affect perceptions of prior disability in a structured interview. *Journal of Business and Psychology*, *20*, 489–500.

Robertson, I. T., Gratton, L., & Rout, U. (1990). The validity of situational interviews for administrative jobs. *Journal of Organizational Behavior*, *11*, 69–76.

Rudloff, A. (2007). *Complete list of behavioral interview questions*. Retrieved from https://law.duke.edu/sites/default/files/career/Complete_List_of_Behavioral_Interview_Questions_and_Answers.pdfca

Salgado, J. F. (2001). Some landmarks of 100 years of scientific personnel selection at the beginning of the new century. *International Journal of Selection and Assessment*, *9*, 3–8.

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*, 262–274.

Schmitt, N., & Ostroff, C. (1986). Operationalizing the "behavioral consistency" approach: Selection test development based on a content-oriented strategy. *Personnel Psychology*, *39*, 91–108.

Schriesheim, C. A., Castro, S. L., & Cogliser, C. C. (1999). Leader–member exchange (LMX) research: A comprehensive review of theory, measurement, and data-analytic practices. *The Leadership Quarterly*, *10*, 63–113.

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

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*, 149–155.

Sprecher, T. B. (1965). Clarifying anchored rating scales based on performance incidents. *ETS Research Bulletin Series*, *1965*, i–14.

Stanford University Human Resources. (2015). *Behavioral interviewing questions*. Retrieved from https://stanford.app.box.com/s/3mpiz26uebbhf1p2g5ct0s4wjxl2wlzg

Taylor, A., & Greve, H. R. (2006). Superman or the fantastic four? Knowledge combination and experience in innovative teams. *Academy of Management Journal*, *49*, 723–740.

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*, 277–294.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*, 83–91.

Tukey, J. W. (1980). We need both exploratory and confirmatory. *The American Statistician*, *34*, 23–25.

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

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

Wang, L., MacCann, C., Zhuang, X., Liu, O. L., & Roberts, R. D. (2009). Assessing teamwork and collaboration in high school students: A multimethod approach. *Canadian Journal of School Psychology, 24*, 108–124.

Weekley, J. A., & Gier, J. A. (1987). Reliability and validity of the situational interview for a sales position. *Journal of Applied Psychology*, *72*, 484–487.

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 157–182). Mahwah, NJ: Erlbaum.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology*, *52*, 372–376.

Wollack, S., Goodale, J. G., Wijting, J. P., & Smith, P. C. (1971). Development of the survey of work values. *Journal of Applied Psychology*, *55*, 331–338.

Zedeck, S., Imparato, N., Krausz, M., & Oleno, T. (1974). Development of behaviorally anchored rating scales as a function of organizational level. *Journal of Applied Psychology*, *59*, 249–252.

Zedeck, S., Kafry, D., & Jacobs, R. (1976). Format and scoring variations in behavioral expectation evaluations. *Organizational Behavior and Human Performance*, *17*, 171–184.

Appendix A

Structured Job Interview Questions for Applied Social Skills

Communication

1 Please tell us about a time when you had to use your listening skills to overcome a communication problem at work. Please provide details about the background of the situation, the behaviors you carried out in response to that situation, and what the outcome was. (Medical University of South Carolina Human Resources, 2015)

2 Please tell us about a time when you had to communicate difficult information or critical feedback to your supervisor at work. How did you give the information/feedback? Please provide details about the background of the situation, the behaviors you carried out in response to that situation, and what the outcome was. (Medical University of South Carolina Human Resources, 2015)

3 Please tell us about a time when you had to explain a complex or technical idea to a nontechnical or unfamiliar audience. How did you go about doing that? Please provide details about the background of the situation, the behaviors you carried out in response to that situation, and what the outcome was.

Leadership

1. Please tell us about a work situation in which you were not the formal leader but tried to assume a leadership role. Please provide details about the background of the situation, the behaviors you carried out in response to that situation, and what the outcome was. (Stanford University Human Resources, 2015)
2. Please tell us about a time when you felt you were able to build motivation in your coworkers or subordinates at work. Please provide details about the background of the situation, the behaviors you carried out in response to that situation, and what the outcome was. (Rudloff, 2007)
3. Please tell us about a time when you have been a member of a group where two of the members did not work well together. How did you handle that situation? Please provide details about the background of the situation, the behaviors you carried out in response to that situation, and what the outcome was. (Rudloff, 2007)

Persuasion and Negotiation

1. Please tell us about a time when you have tried to gain the trust of others you were working with while trying to negotiate with them. Please provide details about the background of the situation, the behaviors you carried out in response to that situation, and what the outcome was. (Stanford University Human Resources, 2015)
2. Please tell us about a time when you had to bargain or compromise with someone you were working with to arrive at a mutually favorable outcome. Please provide details about the background of the situation, the behaviors you carried out in response to that situation, and what the outcome was. (Rudloff, 2007)
3. Please tell us about a time when you had to persuade a person or group of people to accept a proposal of your idea. How did you go about doing this? Please provide details about the background of the situation, the behaviors you carried out in response to that situation, and what the outcome was. (Rudloff, 2007)

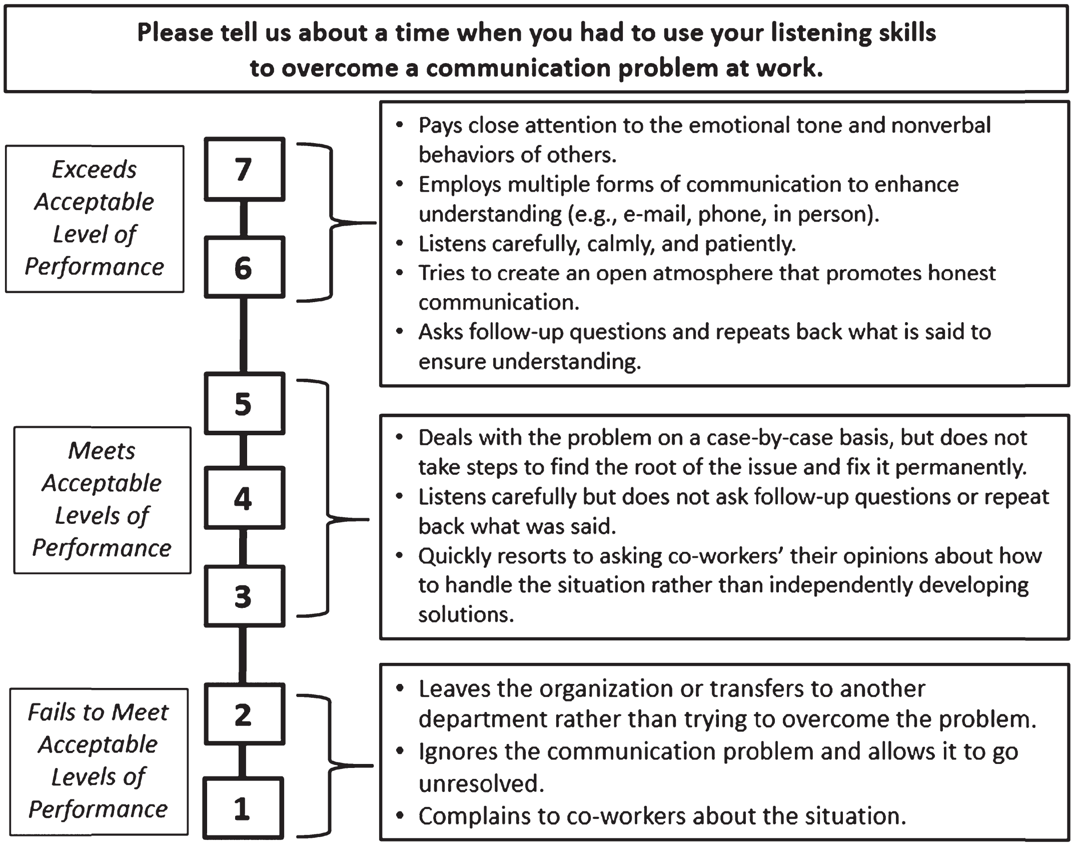
Teamwork

1. Please tell us about a time when you had to work with someone you did not especially like or get along with. How did you interact with this person? Please provide details about the background of the situation, the behaviors you carried out in response to that situation, and what the outcome was.
2. Usually unpleasant tasks (e.g., tedious, boring, physically demanding, etc.) are shared among employees. Please tell us about a time when you thought you were being given more than your share of unpleasant tasks. What did you do? Please provide details about the background of the situation, the behaviors you carried out in response to that situation, and what the outcome was. (Morgeson, Reider, & Campion, 2005)
3. Please tell us about a time when someone took over the leadership of a group project and ignored contributions that were not in accordance with his or her own opinion. How did you handle that situation? Please provide details about the background of the situation, the behaviors you carried out in response to that situation, and what the outcome was. (Klehe & Latham, 2005)

Appendix B

Behavioral Summary Scales

Communication



Please tell us about a time when you had to use your listening skills to overcome a communication problem at work.

Exceeds Acceptable Level of Performance

7

6

- Pays close attention to the emotional tone and nonverbal behaviors of others.
- Employs multiple forms of communication to enhance understanding (e.g., e-mail, phone, in person).
- Listens carefully, calmly, and patiently.
- Tries to create an open atmosphere that promotes honest communication.
- Asks follow-up questions and repeats back what is said to ensure understanding.

Meets Acceptable Levels of Performance

5

4

3

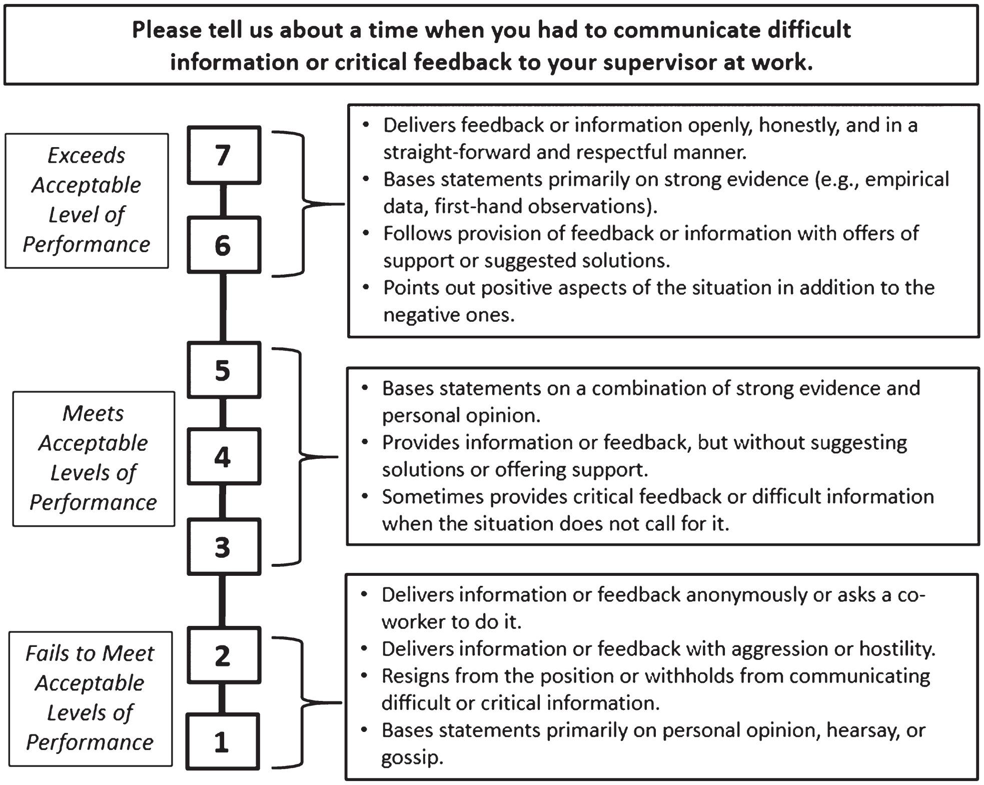
- Deals with the problem on a case-by-case basis, but does not take steps to find the root of the issue and fix it permanently.
- Listens carefully but does not ask follow-up questions or repeat back what was said.
- Quickly resorts to asking co-workers' their opinions about how to handle the situation rather than independently developing solutions.

Fails to Meet Acceptable Levels of Performance

2

1

- Leaves the organization or transfers to another department rather than trying to overcome the problem.
- Ignores the communication problem and allows it to go unresolved.
- Complains to co-workers about the situation.



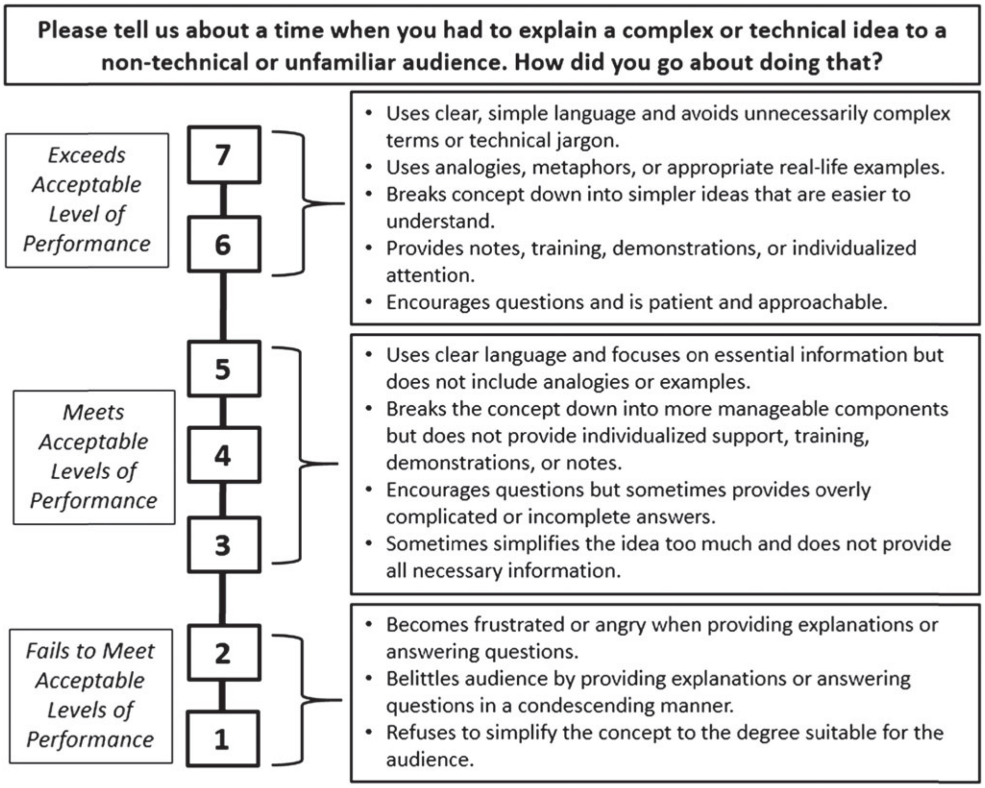
Please tell us about a time when you had to communicate difficult information or critical feedback to your supervisor at work.

Exceeds Acceptable Level of Performance
7
6
Meets Acceptable Levels of Performance
5
4
3
Fails to Meet Acceptable Levels of Performance
2
1

- Delivers feedback or information openly, honestly, and in a straight-forward and respectful manner.
- Bases statements primarily on strong evidence (e.g., empirical data, first-hand observations).
- Follows provision of feedback or information with offers of support or suggested solutions.
- Points out positive aspects of the situation in addition to the negative ones.

- Bases statements on a combination of strong evidence and personal opinion.
- Provides information or feedback, but without suggesting solutions or offering support.
- Sometimes provides critical feedback or difficult information when the situation does not call for it.

- Delivers information or feedback anonymously or asks a co-worker to do it.
- Delivers information or feedback with aggression or hostility.
- Resigns from the position or withholds from communicating difficult or critical information.
- Bases statements primarily on personal opinion, hearsay, or gossip.



Please tell us about a time when you had to explain a complex or technical idea to a non-technical or unfamiliar audience. How did you go about doing that?

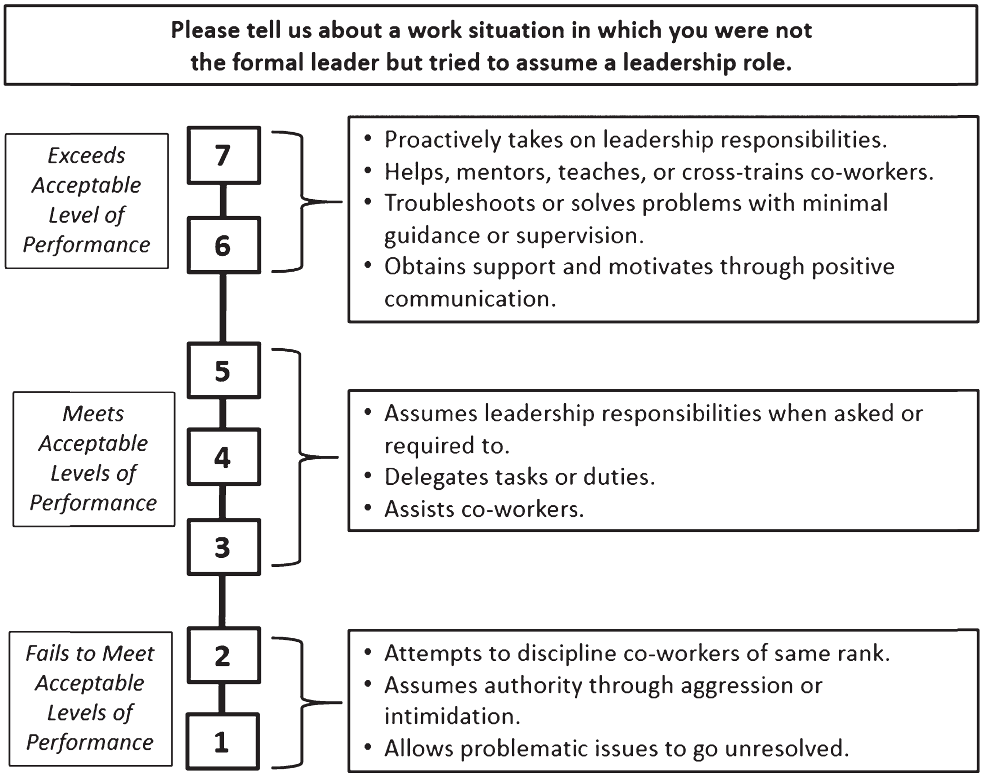
Exceeds Acceptable Level of Performance
7
6
Meets Acceptable Levels of Performance
5
4
3
Fails to Meet Acceptable Levels of Performance
2
1

- Uses clear, simple language and avoids unnecessarily complex terms or technical jargon.
- Uses analogies, metaphors, or appropriate real-life examples.
- Breaks concept down into simpler ideas that are easier to understand.
- Provides notes, training, demonstrations, or individualized attention.
- Encourages questions and is patient and approachable.

- Uses clear language and focuses on essential information but does not include analogies or examples.
- Breaks the concept down into more manageable components but does not provide individualized support, training, demonstrations, or notes.
- Encourages questions but sometimes provides overly complicated or incomplete answers.
- Sometimes simplifies the idea too much and does not provide all necessary information.

- Becomes frustrated or angry when providing explanations or answering questions.
- Belittles audience by providing explanations or answering questions in a condescending manner.
- Refuses to simplify the concept to the degree suitable for the audience.

Leadership



Please tell us about a work situation in which you were not the formal leader but tried to assume a leadership role.

Exceeds Acceptable Level of Performance — 7, 6

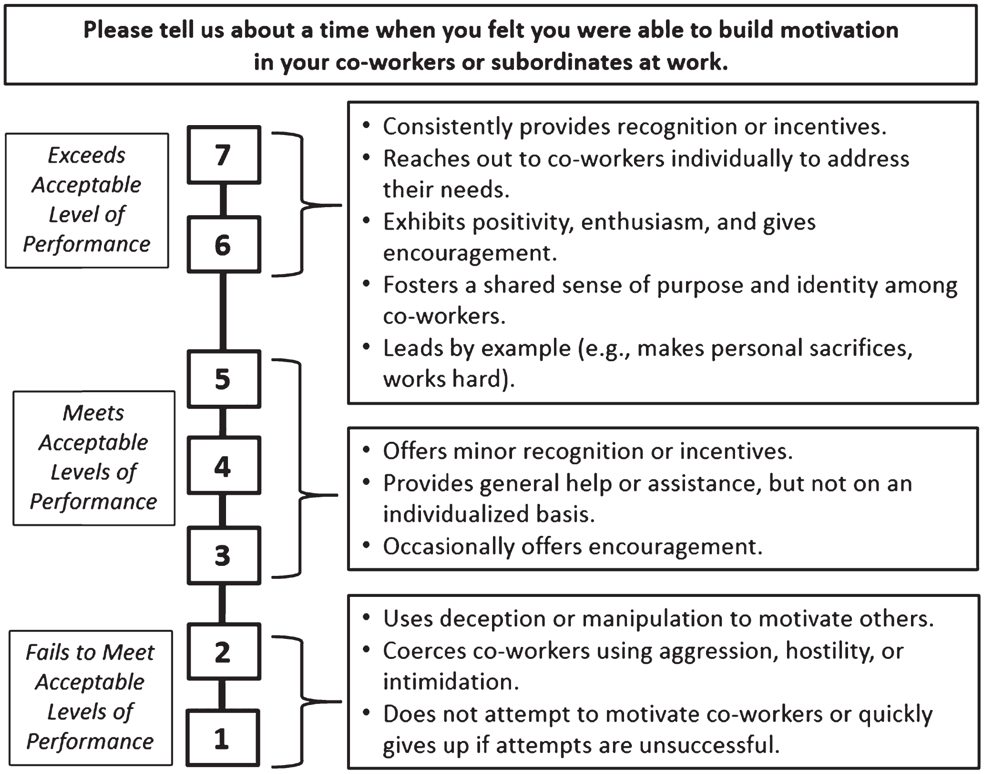
- Proactively takes on leadership responsibilities.
- Helps, mentors, teaches, or cross-trains co-workers.
- Troubleshoots or solves problems with minimal guidance or supervision.
- Obtains support and motivates through positive communication.

Meets Acceptable Levels of Performance — 5, 4, 3

- Assumes leadership responsibilities when asked or required to.
- Delegates tasks or duties.
- Assists co-workers.

Fails to Meet Acceptable Levels of Performance — 2, 1

- Attempts to discipline co-workers of same rank.
- Assumes authority through aggression or intimidation.
- Allows problematic issues to go unresolved.



Please tell us about a time when you felt you were able to build motivation in your co-workers or subordinates at work.

Exceeds Acceptable Level of Performance — 7, 6

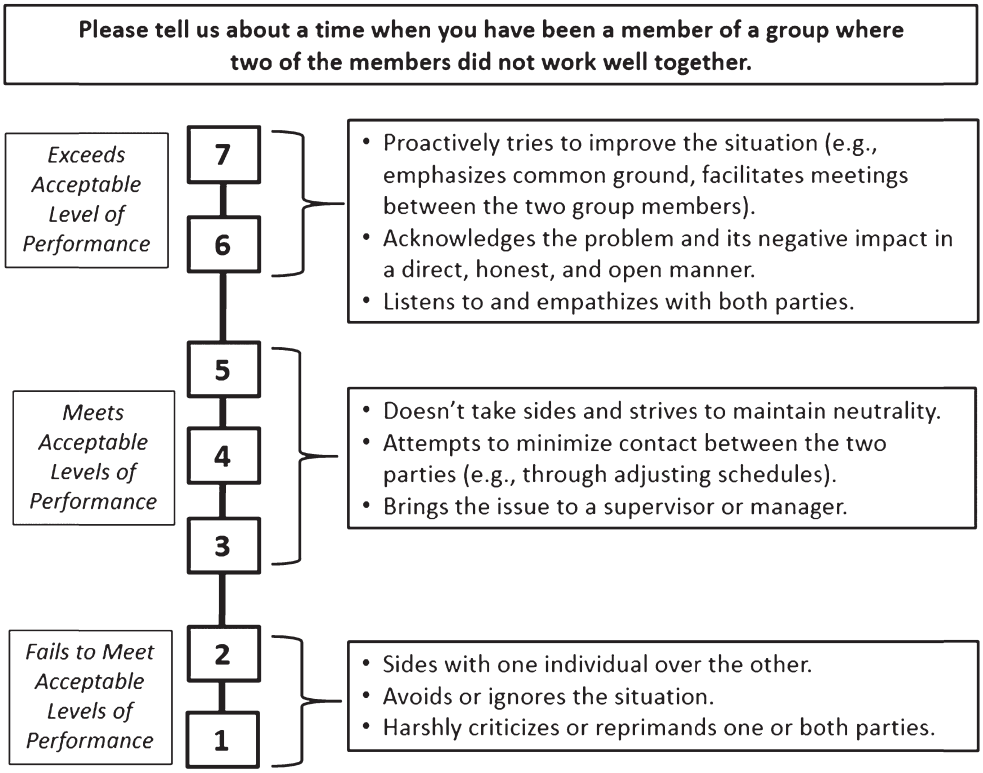
- Consistently provides recognition or incentives.
- Reaches out to co-workers individually to address their needs.
- Exhibits positivity, enthusiasm, and gives encouragement.
- Fosters a shared sense of purpose and identity among co-workers.
- Leads by example (e.g., makes personal sacrifices, works hard).

Meets Acceptable Levels of Performance — 5, 4, 3

- Offers minor recognition or incentives.
- Provides general help or assistance, but not on an individualized basis.
- Occasionally offers encouragement.

Fails to Meet Acceptable Levels of Performance — 2, 1

- Uses deception or manipulation to motivate others.
- Coerces co-workers using aggression, hostility, or intimidation.
- Does not attempt to motivate co-workers or quickly gives up if attempts are unsuccessful.



Please tell us about a time when you have been a member of a group where two of the members did not work well together.

Exceeds Acceptable Level of Performance

7
6

- Proactively tries to improve the situation (e.g., emphasizes common ground, facilitates meetings between the two group members).
- Acknowledges the problem and its negative impact in a direct, honest, and open manner.
- Listens to and empathizes with both parties.

Meets Acceptable Levels of Performance

5
4
3

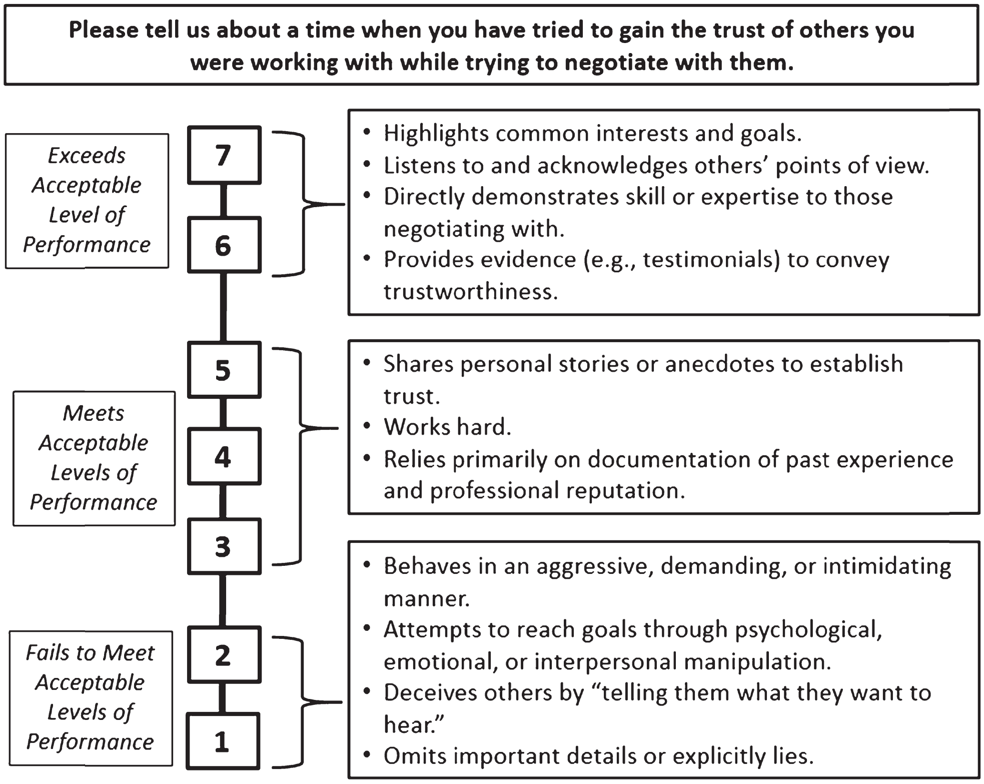
- Doesn't take sides and strives to maintain neutrality.
- Attempts to minimize contact between the two parties (e.g., through adjusting schedules).
- Brings the issue to a supervisor or manager.

Fails to Meet Acceptable Levels of Performance

2
1

- Sides with one individual over the other.
- Avoids or ignores the situation.
- Harshly criticizes or reprimands one or both parties.

Persuasion and Negotiation



Please tell us about a time when you have tried to gain the trust of others you were working with while trying to negotiate with them.

Exceeds Acceptable Level of Performance

7
6

- Highlights common interests and goals.
- Listens to and acknowledges others' points of view.
- Directly demonstrates skill or expertise to those negotiating with.
- Provides evidence (e.g., testimonials) to convey trustworthiness.

Meets Acceptable Levels of Performance

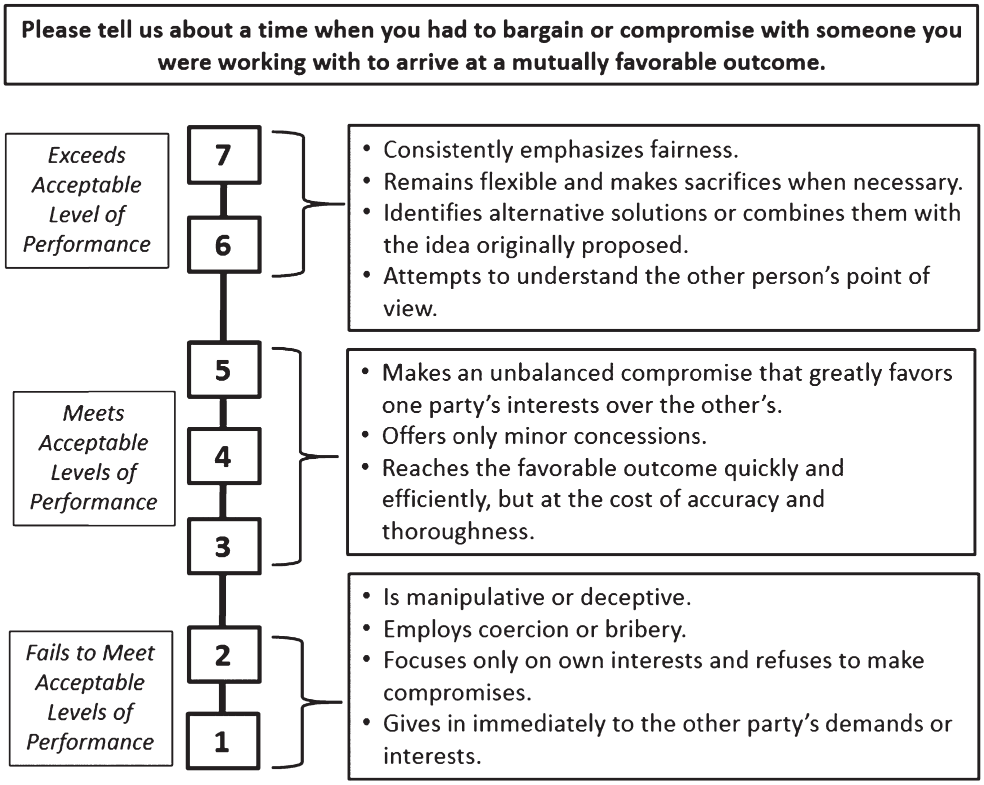
5
4
3

- Shares personal stories or anecdotes to establish trust.
- Works hard.
- Relies primarily on documentation of past experience and professional reputation.

Fails to Meet Acceptable Levels of Performance

2
1

- Behaves in an aggressive, demanding, or intimidating manner.
- Attempts to reach goals through psychological, emotional, or interpersonal manipulation.
- Deceives others by "telling them what they want to hear."
- Omits important details or explicitly lies.

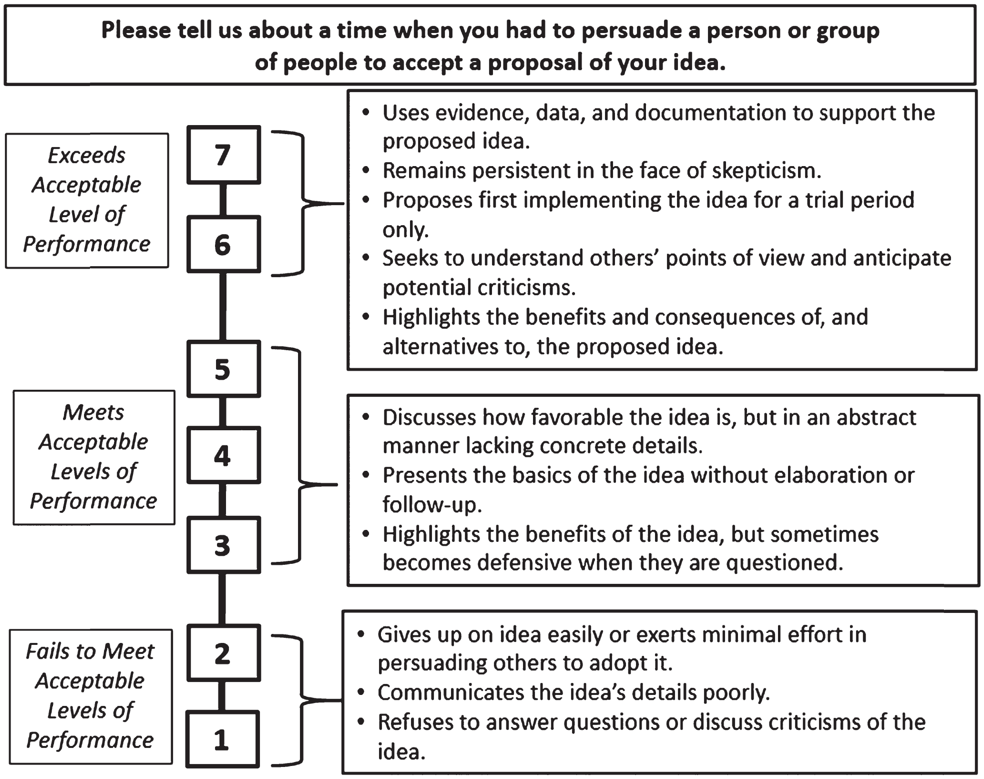


Please tell us about a time when you had to bargain or compromise with someone you were working with to arrive at a mutually favorable outcome.

- Consistently emphasizes fairness.
- Remains flexible and makes sacrifices when necessary.
- Identifies alternative solutions or combines them with the idea originally proposed.
- Attempts to understand the other person's point of view.

- Makes an unbalanced compromise that greatly favors one party's interests over the other's.
- Offers only minor concessions.
- Reaches the favorable outcome quickly and efficiently, but at the cost of accuracy and thoroughness.

- Is manipulative or deceptive.
- Employs coercion or bribery.
- Focuses only on own interests and refuses to make compromises.
- Gives in immediately to the other party's demands or interests.



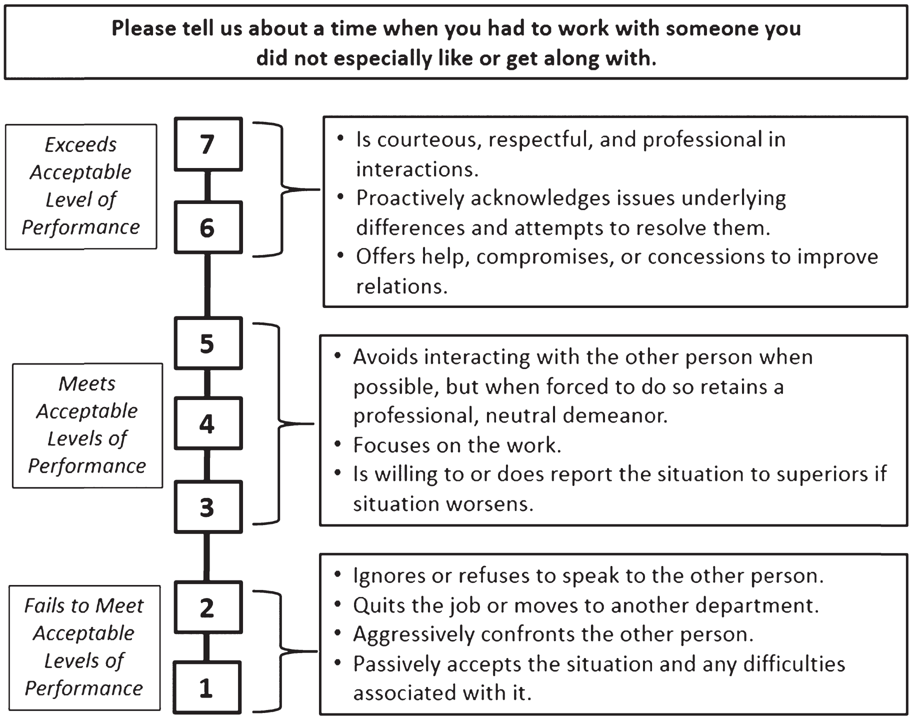
Please tell us about a time when you had to persuade a person or group of people to accept a proposal of your idea.

- Uses evidence, data, and documentation to support the proposed idea.
- Remains persistent in the face of skepticism.
- Proposes first implementing the idea for a trial period only.
- Seeks to understand others' points of view and anticipate potential criticisms.
- Highlights the benefits and consequences of, and alternatives to, the proposed idea.

- Discusses how favorable the idea is, but in an abstract manner lacking concrete details.
- Presents the basics of the idea without elaboration or follow-up.
- Highlights the benefits of the idea, but sometimes becomes defensive when they are questioned.

- Gives up on idea easily or exerts minimal effort in persuading others to adopt it.
- Communicates the idea's details poorly.
- Refuses to answer questions or discuss criticisms of the idea.

Teamwork



Please tell us about a time when you had to work with someone you did not especially like or get along with.

Exceeds Acceptable Level of Performance

7
6

- Is courteous, respectful, and professional in interactions.
- Proactively acknowledges issues underlying differences and attempts to resolve them.
- Offers help, compromises, or concessions to improve relations.

Meets Acceptable Levels of Performance

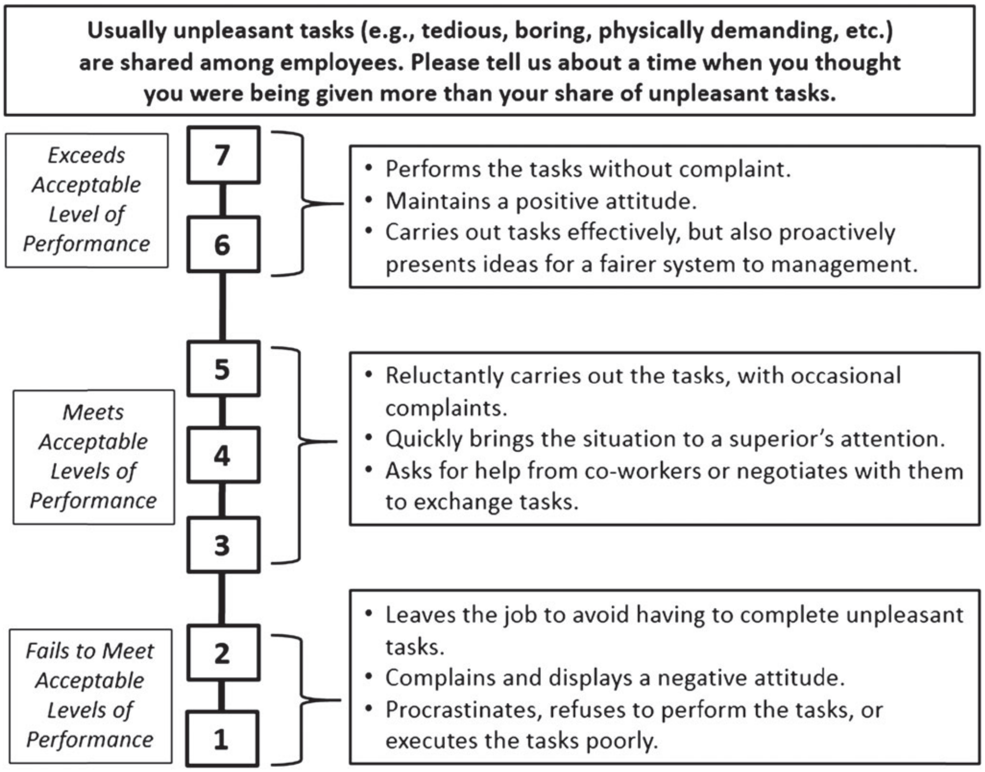
5
4
3

- Avoids interacting with the other person when possible, but when forced to do so retains a professional, neutral demeanor.
- Focuses on the work.
- Is willing to or does report the situation to superiors if situation worsens.

Fails to Meet Acceptable Levels of Performance

2
1

- Ignores or refuses to speak to the other person.
- Quits the job or moves to another department.
- Aggressively confronts the other person.
- Passively accepts the situation and any difficulties associated with it.



Usually unpleasant tasks (e.g., tedious, boring, physically demanding, etc.) are shared among employees. Please tell us about a time when you thought you were being given more than your share of unpleasant tasks.

Exceeds Acceptable Level of Performance

7
6

- Performs the tasks without complaint.
- Maintains a positive attitude.
- Carries out tasks effectively, but also proactively presents ideas for a fairer system to management.

Meets Acceptable Levels of Performance

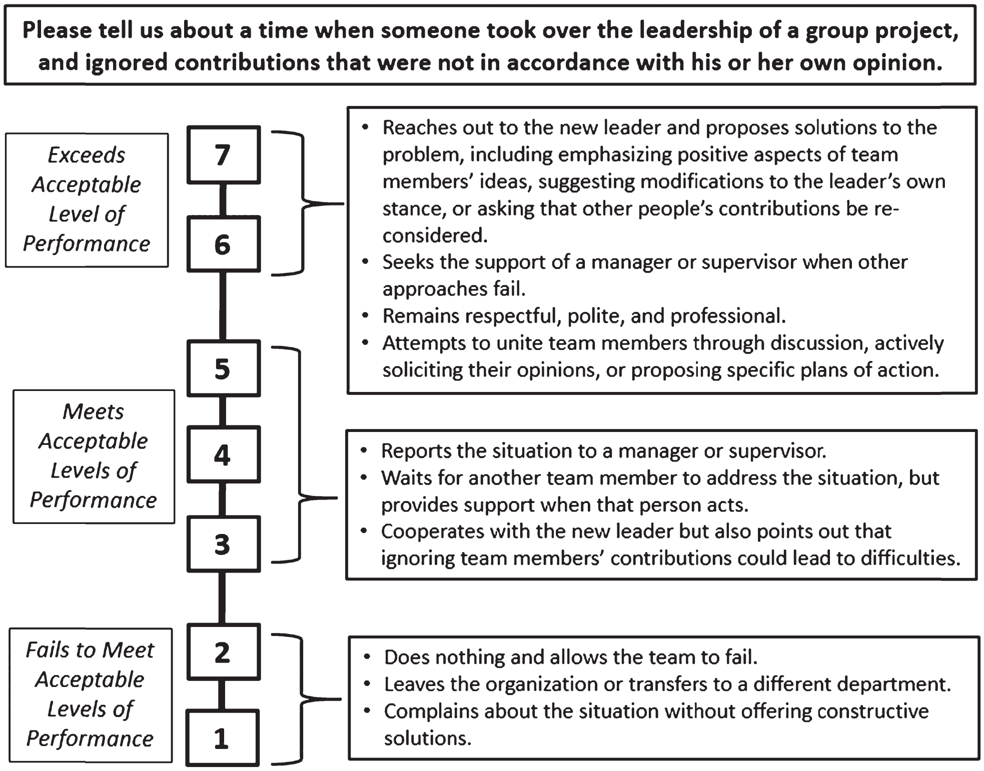
5
4
3

- Reluctantly carries out the tasks, with occasional complaints.
- Quickly brings the situation to a superior’s attention.
- Asks for help from co-workers or negotiates with them to exchange tasks.

Fails to Meet Acceptable Levels of Performance

2
1

- Leaves the job to avoid having to complete unpleasant tasks.
- Complains and displays a negative attitude.
- Procrastinates, refuses to perform the tasks, or executes the tasks poorly.



Please tell us about a time when someone took over the leadership of a group project, and ignored contributions that were not in accordance with his or her own opinion.

Exceeds Acceptable Level of Performance
7
6
Meets Acceptable Levels of Performance
5
4
3
Fails to Meet Acceptable Levels of Performance
2
1

- Reaches out to the new leader and proposes solutions to the problem, including emphasizing positive aspects of team members' ideas, suggesting modifications to the leader's own stance, or asking that other people's contributions be re-considered.
- Seeks the support of a manager or supervisor when other approaches fail.
- Remains respectful, polite, and professional.
- Attempts to unite team members through discussion, actively soliciting their opinions, or proposing specific plans of action.

- Reports the situation to a manager or supervisor.
- Waits for another team member to address the situation, but provides support when that person acts.
- Cooperates with the new leader but also points out that ignoring team members' contributions could lead to difficulties.

- Does nothing and allows the team to fail.
- Leaves the organization or transfers to a different department.
- Complains about the situation without offering constructive solutions.

Suggested citation:

Kell, H. J., Martin-Raugh, M. P., Carney, L. M., Inglese, P.A., Chen, L., & Feng, G. (2017). *Exploring methods for developing behaviorally anchored scales for evaluating structured interview performance* (Research Report No. RR-17-28). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12152

Action Editor: John Sabatini

Reviewers: Sam Rikoon and Larry Stricker

ETS, the ETS logo, and MEASURING THE POWER OF LEARNING.are registered trademarks of Educational Testing Service (ETS). All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/