Applied HRM Research, Volume 11, Number 2, pages 79-96

Puzzle Interviews: What Are They and What Do They Measure?

Jeremiah Honer
California State Personnel Board

Chris W. Wright
San Francisco State University

Chris J. Sablynski
California State University, Sacramento

Recently organizations have begun using unorthodox interviewing techniques and non-traditional interview questions to hire applicants. The most notable example of this trend is the so-called "puzzle interview" which was popularized by Microsoft in the 1990s. Puzzle interviews either ask the applicant to solve puzzles (e.g., "Why are manhole covers round?") or unusual problems (e.g., "How would you weigh an airplane without a scale?"). The goals of the present study were a) to determine if a puzzle interview could be reliably administered and scored, b) to examine whether variation in puzzle interview performance was related to cognitive ability, and c) to explore the relationships between puzzle interview performance and applicant reaction measures. Seventy-six undergraduates participated in a videotaped puzzle interview and completed measures of cognitive ability and applicant reactions. Interview performance ratings by three independent raters yielded an acceptable level of inter-rater reliability. In addition, puzzle interview performance was significantly correlated with cognitive ability and applicant reaction measures of procedural justice and perceptions of performance. Implications for the use of puzzle interviews in industry and for future research are discussed.

The employment interview is one of the most popular selection and recruiting methods (Posthuma, Morgeson, & Campion, 2002). Employers use interviews for a variety of purposes, including initial screening of job applicants, measurement of job-related attributes, and assessment of person-organization fit. There is also considerable variability in how employers structure and administer interviews, what types of questions they ask, and how they score them. Decades of research on the validity of the employment interview show that it can be a useful tool in predicting future job performance, particularly if it is highly structured, designed to ask behavioral or situational questions about job-related characteristics, and scored using behaviorally anchored rating scales (Taylor & Small, 2002).

Recently, however, some organizations have begun using unorthodox interviewing techniques and non-traditional interview questions to hire applicants. The most notable example of this trend is the so-called "puzzle interview". The puzzle interview was popularized by Microsoft in the 1990s, and is now used in other organizations. According to Poundstone (2003):

> *Puzzle-laden job interviews have infiltrated the Fortune 500 and the rust belt; law firms, banks, consulting firms, and the insurance industry; airlines, media, advertising, and even the armed forces. Brainteaser questions are reported from Italy, Russia, and India. Like it or not, puzzles and riddles are a hot new trend in hiring.* (p. 7)

Puzzle interviews are generally administered in an unstructured format and do not use formal scoring keys. The most common types of questions either ask the applicant to solve puzzles or brainteasers (e.g., "Why are manhole covers round?" or "How many times a day do a clock's hands overlap?") or to solve unusual problems (e.g., "How would you weigh an airplane without a scale?" or "How many golf balls does it take to fill an Olympic-sized swimming pool?"). The conventional wisdom among employers is that these questions measure a variety of applicant characteristics, including critical thinking, creativity, intuition, flexibility, and ability to reason (Munk & Oliver, 1997). In our search of the literature, however, we could find no research evidence to support the use of non-traditional questions in the employment interview. Because of this lack of research, the goals of our study were, a) to determine if a puzzle interview could be reliably administered and scored and, b) to examine whether variation in puzzle interview performance could be accounted for by cognitive ability and perceptions of fairness.

Developing the Puzzle Interview

That we know of, there is no previous research that has examined the psychometric properties of puzzle interviews. Before testing hypotheses related to the usefulness of such interviews, their psychometric soundness must first be established. We attempted to do this by a) developing anchored rating scales to use as a guideline for scoring responses, and b) using a panel of raters to score the responses. The goal was to ascertain the level of inter-rater reliability in scoring. If a reasonable degree of inter-item reliability emerged, we would assume that the questions were being measured consistently, and research could continue to test subsequent hypotheses in the study.

We followed the process established by Campion, Pursell, and Brown (1988) who substantially increased the reliability and validity of structured interviews in a field study using a five-step model. First, questions were based on a job analysis. Second, the same questions were asked of each candidate. Third, the ratings scales were anchored to score answers with examples and illustrations. Fourth, an interview panel was used to record and rate answers. Finally, the process was administered to all candidates consistently. This included not discussing questions or answers with other panel members between interviews.

With one exception, we used this process to design the puzzle interview. Because the study was designed to assess the general appropriateness of puzzle interviews, no job analysis was conducted; and the questions were not linked to job analysis information. However, the puzzle interview was constructed to include the most commonly used types of questions (Poundstone, 2003). For the purpose of this

study, those questions were categorized into two different types, *proscriptive* and *descriptive*.

Proscriptive problems elicit the applicant's problem solving ability and have an identifiable correct answer. An example of a proscriptive problem would be to ask the applicant to measure exactly four gallons of water using only a 3-gallon and a 5-gallon jug. Descriptive problems, on the other hand, have no single correct answer. Interviewers believe descriptive questions assess the applicant's creativity, resourcefulness, or ability to think through a problem by factoring in numerous variables (Poundstone, 2003). An example of a descriptive interview problem would be to ask the applicant to estimate how many gas stations there are in the United States. A superior answer would identify relevant variables (such as the number of licensed drivers, gas mileage of vehicles, average traveling distance of an individual) and describe a process for using them to determine a solution to the problem.

Puzzle Interviews and Cognitive Ability

As indicated above, employers assume that puzzle interviews are associated with cognitive ability, but there is no scientific literature to support this assertion (Monk & Oliver, 1997; Poundstone, 2003). However, traditional interview formats have been directly linked to cognitive ability in several studies. Campion, Campion, and Hudson (1994), for example, found that a 30-item structured interview containing equal amounts of behavioral and situational questions was significantly related to a composite of two cognitive job performance dimensions, learning orientation and technical knowledge ($r = .60$, $p < .01$). In addition, Huffcutt, Roth, and McDaniel (1996) used meta-analytic data to show that, on average, cognitive ability accounts for 16 percent of the variance in employment interview evaluations. They also found that situational interview questions correlated more highly with cognitive ability than behavior description interview questions. This finding is consistent with that of McDaniel, Whetzel, Schmidt, and Maurer (1994). Situational interviews are thought to load more heavily on verbal and inductive reasoning abilities than behavior description questions (Janz, 1989).

Puzzle interview questions are somewhat similar to situational interview questions in that a majority of them present interviewees with novel situations. The nature of the questions, however, is entirely different from those used in a traditional situational interview. Puzzle questions more closely resemble written questions of problem solving ability, which have also been linked to cognitive ability (Raaheim & Brun, 1985; Vernon & Strudensky, 1988).

In one study, for example, Vernon and Strudensky (1988) found a modest relationship between intelligence and the solution of two problems that involved completing a task with a clearly defined set of rules. One of the problems was the classic Missionaries-Cannibals dilemma, in which the participant was required to lead three missionaries and three cannibals across a river using a boat for two. The constraint was that the cannibals could not outnumber the missionaries on either side of the river at anytime. Factor analysis of responses to the problem yielded a general intelligence dimension. A 'move' and a 'speed' factor were also identified,

indicating that participants scoring higher on an IQ battery made fewer moves and completed the puzzle in a shorter time.

Raaheim and Brun (1985) also found a relationship between cognitive ability and problem solving. In their study, participants with a pre-existing knowledge of geometry were asked to determine the number of calories in a piece of cake cut out from a different geometric shape. The findings suggested a relationship between intelligence and problem solving ability. The results are supported by research indicating that other problems of this type correlate with scores on general mental ability tests (Mackintosh, 1998).

Based on the research relating intelligence to problem solving and the established relationship between cognitive ability and structured interview performance, the following hypothesis was tested in the current study:

> *Hypothesis 1: Cognitive ability will be significantly correlated with performance on the puzzle interview, such that higher cognitive ability will be associated with higher interview performance.*

Applicant Reactions to Selection Procedures

Hausknecht, Day, and Thomas (2004) cite five reasons why applicant reactions to such unorthodox selection procedures as puzzle interviews are important for employers. First, the effects that applicant reactions have on organizational attractiveness may indirectly influence the pursuit of job offers with the organization (Murphy, 1986). Second, applicants who have a negative view of the selection process may be less likely to accept a formal job offer (Macan, Avedon, Paese, & Smith, 1994). Third, negative experiences with an organization's selection procedure may be more likely to persuade other applicants from pursuing employment with the organization (Smither, Reilly, Millsap, Pearlman & Stoffey, 1993). Fourth, negative applicant reactions may be related to both the likelihood of litigation and to the successful defense of the selection procedure (Smither, et al. 1993). Fifth, applicants who have a negative reaction to a selection procedure may be more likely to avoid engaging in consumer behavior with the organization in the future. Rynes and Barber (1990) referred to this as the "spillover effect." An additional consequence of negative applicant reactions, offered by Smither et al. (1993), is that such reactions may indirectly affect both the validity and utility of the selection instrument by impacting variables such as test taker motivation.

There is considerable research examining perceptions of fairness of selection instruments and the question of what features of a selection method increase the applicant's perceptions of fairness (Gilliland, 1993; Lounsbury, Bobrow, & Jensen, 1989). For example, research has suggested that selection procedures are perceived more fairly if applicants view them as more concrete and job related. Smither, et al. (1993) gathered responses of perceived fairness (defined as perceived face validity and perceived predictive validity) from subjects who viewed a selection process. Simulations, interviews, and cognitive ability tests with more concrete item types (i.e. standard written English, vocabulary in context, mathematical word problems) were perceived as having greater validity than personality inventories, biodata and

cognitive ability tests with more abstract items (e.g., letter sets, quantitative comparisons). In a similar study, Kluger and Rothstein (1993) found that participants who assumed that they were assisting in the computerization of a selection procedure for a local employer had more positive reactions to a biographical inventory than an abstract cognitive ability test.

Two additional variables that influence participant perception of fairness of a selection procedure are actual performance and perceived performance. For example, Chan, Schmitt, Sacco, and DeShon (1998) offered participants a monetary incentive for high performance on a cognitive ability test. Participants viewed sample test items and gave pre-test reaction data. They then took the full test and completed post-test reaction measures. Correlations between pretest and posttest reactions were significant, and posttest reactions correlated with performance, even after partialling out pretest reactions. This supports the hypothesis that posttest reactions are partly a function of test performance (Chan et al., 1998).

It has also been suggested that actual performance ratings may be deficient in determining a relationship between performance and perceptions of fairness, because applicants may be inaccurate in determining their performance (Chan et al., 1998). Macan, Avedon, Paese, and Smith (1994) addressed this issue using perceived performance in a study of applicants' perceptions of content and predictive validity for various test types. Participants voluntarily completed a questionnaire after taking a cognitive ability test and ratings of perceived performance accounted for 16 percent of the variance in perceptions of fairness ratings (Macan et al., 1994).

Because previous research has found relationships between perceptions of fairness, actual performance, and perceived performance, the following hypotheses were tested in this study:

> *Hypothesis 2: Perceptions of fairness will be significantly correlated with performance on the puzzle interview, such that higher performance will be associated with higher perceptions of fairness.*
>
> *Hypothesis 3: Perceptions of fairness will be significantly correlated with perceived performance on the puzzle interview, such that higher levels of perceived performance will be associated with higher perceptions of fairness.*

Method

Participants

A sample of 76 participants was obtained from psychology undergraduate courses at a university in the western United States in the spring of 2005. The participants consisted of 24 men and 52 women, with a mean age of 25.26 years (SD = 7.94). The ethnic composition of the sample was 27.6% Caucasian, 7.9% African-American/Black, 27.6%, Asian, 19.7% Hispanic/Latino, and 17.1% "Other". Most of the participants (77%) were employed (mean hours per week = 26.17, SD = 10.34). In addition, a majority of participants (65%) reported that their level of

interviewing experience was "Somewhat experienced" or greater. Although the ideal sample would have included actual job applicants, Wiesner, Saks, and Summers (1991) have suggested that the characteristics of the undergraduate student population are comparable to recent graduates seeking entry-level jobs.

Procedure

Puzzle Interview Development. Items were selected from the popular text by William Poundstone (2003) and were chosen to be representative of the types of puzzle questions that are asked frequently in organizational settings. All interview items were scored using a scale ranging from 1 "Poor" to 5 "Excellent." Anchored rating scales were developed so that a score of "1" typically represented non-response or an inability to understand the question, whereas a score of "5" demonstrated clear understanding of the problem and the ability to describe the answer within the constraints provided.

Puzzle Interview Administration. The primary investigator and two research assistants administered and scored the interviews. The research assistants were trained in structured interview administration and scoring prior to data collection. They were trained using a modified frame-of-reference procedure that involved viewing pilot videos and providing interview scores for participants. Scoring was then discussed to ensure that the anchored rating scale was interpreted similarly by all raters. The research assistants practiced interview administration by interviewing volunteers and receiving performance feedback from the primary investigator. Structured probes were provided to ensure consistent interview administration across all interviewers.

Participants for this study were interviewed in a laboratory setting. All interviews were videotaped to enable scoring by raters at a later time. Prior to the interview, participants were given a written set of instructions, informing them of the interviewing procedures. The same instructions were read aloud to all participants before the reading the first interview question. The instructions were read as follows:

> *Interviews such as the one you are about to participate in have become a popular part of the selection process in many industries. Questions may involve analytical problem solving or creative thought. Your task is to imagine that you are interviewing for an entry-level position that requires some degree of analytical problem solving. You will be asked a series of interview questions and prompted to answer. Remember that you are in a mock interview and that it is in your interest to impress the interviewer. Answer all questions as thoughtfully and thoroughly as you can. As appreciation for your participation, gift certificates to a major retail establishment will be given to the top five performers. The questions are very difficult. Try to relax and do the best that you can. You have five questions and 30 minutes. Manage your time well.*

Participants were given the opportunity to ask questions about the puzzle interview before its initiation. Unlimited scratch paper was supplied for participants to use at their discretion, though they were told to verbalize their responses. During

the interview, the following five questions were read to the participants in the order listed:

1. You have a 3-quart bucket, a 5-quart bucket, and an infinite supply of water. How can you measure out exactly 4 quarts?
2. If the United States was interested in reducing the number of states from 50 to 49, which state should be eliminated and why?
3. How many gas stations are in the United States? Please give an estimate, and describe how you arrived at your estimate.
4. You have eight billiard balls. One of them is "defective," meaning that it weighs more than the others. How do you tell, using a balance, which ball is defective in two weighings?
5. You have a bucket of jellybeans in three colors – red, green, and blue. With your eyes closed, you have to reach in the bucket and take out jellybeans. What is the minimum number of jellybeans you can take out to be certain of getting two of the same color?

Interviewers' interactions with participants were limited to: a) rereading the question, b) confirming information that was contained in the question, and c) using one of the following probes to elicit measurable responses from participants:

1. Please try and think aloud, we would like to record your thought process.
2. Can you describe how you're going about solving the problem?
3. Can you describe how you came up with that answer/estimation/choice?
4. It's important that you try to answer the questions.
5. Is there anything else you'd like to add?

Following the interview, participants completed several measures:

Measures

Perceptions of Fairness. Scales developed by Smither et al. (1993) to measure a participant's perceptions of fairness of personnel selection instruments were modified slightly so the wording referred directly to the puzzle interview. Perceived predictive validity was measured using five items ($\alpha = .84$) that measured the degree to which participants believed the interview indicated how well a person would perform on a job. An example item from this scale was, "The employer can tell a lot about the applicant's ability to do the job from the results of the interview." Five items measured perceived face validity ($\alpha = .79$), providing a general index of how participants perceived the interview to be related to the job. An example item from this scale was, "The actual content of the interview was clearly related to the job." Procedural justice was measured with a two-item scale ($\alpha = .69$). An example item was, "I felt good about the way the interview was conducted and administered." All scales mentioned in this section used a 7-point Likert scale ranging from 1 = "strongly disagree" to 7 = "strongly agree."

Perceived Performance. Four items were used to determine how well the participants believed they performed on the puzzle interview ($\alpha = .89$). Two items

measured how well participants believed they performed in the interview in general terms (e.g., "I did well on this interview"), while two items measured how well participants believed they performed in relation to other participants (e.g., "Compared to other applicants, I did well on the interview").

Cognitive ability. Cognitive ability was measured using the Wonderlic Personnel Test (WPT). Test-retest reliabilities for the WPT have ranged from .82 to .94. Internal consistency has ranged from .88 to .94 (Wonderlic, 2002). Its correlations with instruments such as the Wechsler Adult Intelligence Scale (WAIS) and the General Aptitude Test Battery's "Aptitude G" (for general mental ability or intelligence) range from .70-.92.

Results

Puzzle Interview Reliability

Three independent raters scored the interview responses for each participant. The scoring key for the questions is in Appendix A. The raters were blind to all data collected (i.e., cognitive ability and fairness perceptions) when interview scoring took place. All five questions included in the puzzle interview demonstrated reasonable inter-rater reliability. Intra-class correlations for the each item are included in Table 1, along with the means and standard deviations. Internal consistency of the interview was adequate for a five-item scale ($\alpha = .62$) and was consistent with prior meta-analytic findings on the internal consistency of structured employment interviews (Conway, Jako, & Goodman, 1995). The authors concluded that the levels of inter-rater reliability and internal consistency were acceptable for using puzzle interview scores to explore the study's hypotheses.

Table 1
Descriptive Statistics and Inter-Rater Reliabilities (ICCs) for the Puzzle Interview.

Puzzle Interview	Mean	SD	ICC
Question 1	2.20	1.18	.93
Question 2	2.06	.56	.75
Question 3	1.93	.65	.83
Question 4	2.52	1.14	.95
Question 5	2.64	1.59	.95
Total SPI	11.36	3.44	--

Note: Each question was scored on a Likert scale ranging from 1-5.

N = 76

Table 2
Descriptive Statistics for Perceived Performance Measure, Perceptions of Fairness Measures, and Wonderlic Personnel Test

	Highest Score Possible	M	SD
Perceived Predictive Validity	35	16.75	6.71
Face Validity	35	16.30	6.73
Procedural Justice	14	8.70	3.22
Perceived Performance	26	11.74	4.77
Wonderlic Personnel Test	50	21.73	6.47

Table 2 summarizes the descriptive data for the interview reaction survey that was administered to participants immediately after the interview.

Test of Hypotheses

Pearson correlations were used to test all hypotheses for this study. The first hypothesis stated that cognitive ability would be significantly related to performance on the puzzle interview. This hypothesis was supported, with scores on the Wonderlic Personnel Test accounting for about 20% of the variance in puzzle interview performance ($r = .45$, $p < .01$).

The second hypothesis examined fairness perceptions in relation to actual performance, whereas the third hypothesis examined fairness perceptions in relation to perceived performance. These hypotheses were partially supported. Both actual performance and perceived performance on the puzzle interview were significantly correlated with participant ratings of procedural justice but not to predictive validity or face validity perceptions. The results of these analyses are presented in Table 3.

Table 3
Correlations Between Puzzle Interview Scores and Applicant Reaction Variables (N=76)

Variable	1	2	3	4	5	6
1. Interview score						
2. Perceived interview performance	.31**					
3. Predictive validity	- .20	.04				
4. Face validity	.02	.08	.55**			
5. Procedural justice	.27*	.52**	.25**	.34**		
6. Cognitive ability	.45**	.30**	- .35**	- .07	.08	
7. Sex (0=female, 1=male)	.14	.48**	- .08	- .18	.18	.18

*p < .05, **p < .01

Discussion

This study demonstrated that puzzle interviews may be administered and scored in a structured format that yields acceptable levels of reliability. There is also preliminary evidence to indicate that these interviews may measure cognitive ability. The moderate correlation indicates, however, that a substantial proportion of the variance in puzzle interview performance is accounted for by other constructs. Regarding applicant reactions to puzzle interviews, the findings of this study were mixed. Compared to their counterparts, participants who performed well on the puzzle interview, and those who perceived that they performed well, reported that the process was more procedurally fair. There were no significant differences, however, in ratings of face validity or perceived predictive validity. This may have been due, in part, to the fact that ratings on these two constructs were relatively low for all participants. Means on both the face validity and perceived predictive validity scales were below the midpoint of the maximum scale value. Another explanation for the mixed findings is that it may have been difficult for participants to accurately assess face validity and perceived predictive validity given the fact that they were not given a job description, nor was the interview targeted for a specific position.

Limitations and Future Directions

The primary limitations of this study are related to its external validity. Participants were college students who took part in the study in order to earn academic credit, and there was no penalty for poor performance (i.e. getting hired vs. not getting hired). Thus, our data only reflect how participants believe they would react to a puzzle interview if they were applying for an actual job. The lack of external validity may have affected not only the applicant reactions to the interview, but also interview performance. It is possible that participants would have attempted to solve the problems more thoroughly if they were engaged in an interview for a job that was appealing to them. The researchers attempted to ameliorate this limitation by providing an incentive to perform, in the form of a prize for the five highest scores on the puzzle interview.

Future research should attempt to improve the reliability of the puzzle interview. Though the internal consistency of the puzzle interview was adequate for a five-item scale, there is certainly room for improvement in developing an interview that is reliable enough to allow researchers to conduct meaningful construct and criterion validity studies. Research also needs to be conducted to determine if different types of puzzle interview questions are more reliable than others. This study attempted to categorize the questions into two different question types: proscriptive questions, which have a correct answer, and descriptive questions, which do not. It is possible that internal consistency could be improved by focusing on a single question type.

One of the biggest challenges of this study was the development of scoring guidelines for the questions. Some questions lend themselves to an anchored rating scale format more than others. For example, ratings for the fifth puzzle interview question ("You have a bucket of jelly beans in three colors…") tended to fall at

either the highest or lowest point on the rating scale. This was not the original intention of the authors but occurred because the correct answer to the question requires the application of arithmetic. Participants who recognized this tended to answer the question correctly and thus received the highest scale rating ("5"). However, participants who had difficulty answering the question correctly typically focused their answer on probability estimation, which never led to the correct solution and resulted in the lowest scale rating ("1"). The fourth question ("You have eight billiard balls. One of them is defective…), on the other hand, was well suited for the rating format. There was a clearly defined correct answer, but many participants earned points for offering partial solutions to the problem and the full range of scale values were utilized by the raters. Researchers who attempt to design anchored rating scales for puzzle interview questions in the future should conduct extensive pilot testing in order to avoid this problem.

Research is also needed to examine additional correlates of puzzle interview performance. The correlation between puzzle interview scores and cognitive ability was higher in magnitude than the correlation between cognitive ability and traditional behavioral interviews (Huffcutt et al., 1996). A large percentage of the variance in puzzle interview performance, however, is still unexplained.

Other constructs that may correlate with puzzle interview performance include stress tolerance, critical thinking ability, flexibility, discipline, creativity, and personality characteristics. Affinity for brainteaser problems and/or past experience with brainteaser questions may also predict performance. Additionally, performance differences related to demographic characteristics, academic discipline, and job type/industry should be explored. Regarding job type, puzzle interviews may be more, or less, predictive depending on the complexity of the job. It has already been established that the predictive validity of cognitive ability increases as job complexity increases (Hunter & Hunter, 1984). This may be true for puzzle interviews as well.

More research is also needed to clarify applicant reactions to puzzle interviews. They are often used to assess an interviewee's reactions to a novel, uncomfortable situation. Some employers feel that the widespread availability of information about conventional interviews, through the Internet and other sources, allows candidates to prepare their responses in advance of the formal interview (Poundstone, 2003). Thus the puzzle interview is seen as an opportunity to assess the "real" candidate. It seems logical to assume, however, that job applicants might react negatively to puzzle interviews and that those negative reactions might, in the long term, counteract many of the advantages of using such a procedure. Data gathered in this study suggest that this might be the case, but more sophisticated research on the topic is needed.

Another issue that should be investigated is potential adverse impact. The sample size in this study was not large enough to examine racial subgroup score differences in puzzle interview performance. Because our data suggest that puzzle interview scores may be correlated with cognitive ability, additional research should be conducted to determine if puzzle interviews produce subgroup score differences that would result in adverse impact in selection decisions. If there are subgroup differences, the magnitude of those differences should be compared those of paper-

and-pencil cognitive ability measures (Hunter & Hunter, 1984; Pulakos & Schmitt, 1996).

Subsequent research in this area should also provide significant incentives to enhance applicant motivation. Future laboratory experiments of applicant reactions should also be designed so that applicants believe they are applying for a job in their field of interest. For example, software designers might consider the interview to be more valid than a person who is interested in social work.

The recruitment of high quality employees is critical to the success of any organization, and it is clear that employment interviews will continue to be utilized by organizations to help them achieve this goal. Ultimately then, the most important question for organizations is whether puzzle interviews are an appropriate method for selecting job applicants. The results of this study indicate that these unorthodox interviews may be related to constructs that predict job performance, but there are many more questions that should be addressed before organizations formally adopt these interviews into their selection processes. For example, our results indicate that puzzle interviews are not pure measures of cognitive ability. Employers who wish to measure general intelligence would be advised to use a paper-and-pencil measure of cognitive ability that has been construct-validated. In addition, our findings suggest that administering a puzzle interview to applicants could have a negative impact on their perception of the organization's procedural fairness, which may, in turn, reduce the organization's attractiveness to these prospective employees. Thus, until additional research is conducted, including a formal study in an organizational setting, we recommend that organizations use puzzle interviews with caution.

References

Campion, M.A., Campion, J.E., & Hudson, J. P. (1994). Structured interviewing: A note on incremental validity and alternative question types. *Journal of Applied Psychology, 79*, 998-1102.

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.

Chan, D., Schmitt, N., Sacco, J.M., & DeShon, R.P. (1998). Understanding pretest and posttest reactions to cognitive ability and personality tests. *Journal of Applied Psychology, 83*, 471-485.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80*, 565-579.

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18*, 694-734.

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.

Huffcutt, A.I., Roth, P.L., & McDaniel, M.A. (1996). A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity. *Journal of Applied Psychology, 81*, 459-473.

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

Janz, T. (1989). The patterned behavior description interview: The best prophet of the future is the past. *In* R.W. Eder & G.R. Ferris (Eds.), *The Employment interview: Theory, research, and practice* (158-168). Newbury Park, CA: Sage.

Kluger, A.N., & Rothstein, H.R. (1993). The influence of selection test type on applicant reactions to employment testing. *Journal of Business & Psychology, 8*, 3-25.

Lounsbury, J.W., Bobrow, W., & Jensen, J.B. (1989). Attitudes toward employment testing: Scale development, correlates, and "known-group" validation. *Professional Psychology: Research & Practice, 20*, 340-349.

Macan, T.H., Avedon, M.J. & Paese, M., & Smith, D.E. (1994). The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology, l 47*, 715-738.

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.

Mackintosh, N.J. (1998). IQ and Human Intelligence. New York: Oxford University Press.

Munk, N., & Oliver, S. (1997). Think Fast! *Forbes, 159*(6), 146-150.

Murphy, K.R. (1986). When your top choice turns you down: Effect of rejected job offers on the utility of selection tests. *Psychological Bulletin, 99*, 133-138.

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

Poundstone, W. (2003). How would you move Mount Fuji?: Microsoft's cult of the puzzle. New York: Library of Congress.

Pulakos, E.D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. Human Performance, 9, 241-258.

Raaheim, K., & Brun, W. (1985). Task novelty and intelligence. *Scandinavian Journal of Psychology, 26*, 35-41.

Rynes, S. L., & Barber, A. E. (1990). Applicant attraction strategies: An organizational perspective. *Academy of Management Review*, *15*(2), 286-310.

Smither, J.W., Reilly, R.R., Millsap, R.E., Pearlman, K., & Stoffey, R.W. (1993). Applicant reactions to selection procedures. *Personnel Psychology, 46*, 49-76.

Taylor, P., & 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 & Organizational Psychology, 75*(3), 277-294.

Vernon, P.A., & Strudensky, S. (1988). Relationships between problem-solving and intelligence. *Intelligence, 12*, 435-453.

Wiesner, W. H., Saks, A.M., & Summers, R.J. (1991). Job alternatives and job choice. *Journal of Vocational Behavior, 38*, 198-207.

Wonderlic Personnel Test & Scholastic Level Exam User's Manual. (2002). Libertyville, IL: Wonderlic.

Author Information

Jeremiah Honer
Test Validation and Construction Unit
California State Personnel Board
801 Capitol Mall
Sacramento, CA 94814
Phone: (916) 654-8538
Email: jhoner@spb.ca.gov

Chris W. Wright
Department of Psychology
San Francisco State University
1600 Holloway Avenue
San Francisco, CA 94132
Phone: (415) 338-1465
Email: cwright@sfsu.edu

Chris J. Sablynski
College of Business Administration
California State University, Sacramento
6000 J Street
Sacramento, CA 95819
Phone: (916) 278-7164
E-mail: sablynsk@csus.edu

APPENDIX A

Structured Puzzle Interview (SPI) Scoring Key

1. You have a 3-quart bucket, a 5-quart bucket, and an infinite supply of water. How can you measure out exactly 4 quarts?				
1	2	3	4	5
• Unable to grasp the problem • Suggests to estimate the amount		• Attempts a precise measurement • Suggests pouring amounts from one bottle to another etc. • Describes mathematics and/or correctly states result amount for given weighing		• Provides the correct answer (Fill 3-quart bucket twice; Each time, pour contents into 5-quart bucket; There will be 1 quart left in 3-quart bucket; Empty 5-quart bucket; Pour 1 quart into 5-quart bucket; Fill 3-quart bucket and pour into 5-quart bucket) • Makes very few errors

2. If the United States was interested in reducing the number of states from 50 to 49, which state should be eliminated and why?				
1	2	3	4	5
• Unable to grasp the problem • Gives answer based on preference, with little or no consideration of consequences		• Considers different states. • Discusses some potential consequences: Examples: • Population-i.e. the U.S. would lose less people if this state were removed • What resources would be lost • Impact on the economy • Choosing a non-contiguous state vs. a contiguous one • Demographic/cultural issues		• Considers many consequences • Methodically eliminates states based on a well thought out rationale

3. How many gas stations are in the United States? Please give an estimation and describe how you arrived at your estimate.				
1	2	3	4	5
• Unable to grasp the problem • Gives an estimate with no rationale		• Develops/discusses some rationale for estimation Examples: • Number of cars • Number of people, and how many people have cars • Number of licensed drivers • How often each car is "filled up" • Amount of gas used in a time period • Average distance traveled • Average gas mileage		• Develops a specific rationale, linking numerous variables • Example: estimating the number of cars by relating it to the U.S. population, accounting for family size and how many cars there are per person

4. You have eight billiard balls. One of them is "defective," meaning that it weighs more than the others. How do you tell, using a balance, which ball is defective in two weighings?				
1	2	3	4	5
• Unable to grasp the problem • Takes a guess and provides no rationale		• Attempts various methods. • Is able to correctly vocalize one step to the next, remaining within the constraints of the problem.		• Provides the correct answer (Place any three balls on each side of the scale; If the scale balances, then one of the two remaining balls is defective; If the scale does not balance, one of the three balls from the lower side is defective; For either scenario, use the second weighing to determine which ball is defective) • Makes very few errors

5. You have a bucket of jellybeans in three colors – red, green, and blue. With your eyes closed, you have to reach in the bucket and take out jellybeans. What is the minimum number of jellybeans you can take out to be certain of getting two of the same color?				
1	2	3	4	5
• Unable to grasp the problem • Takes a guess and provides no rationale		• Discusses possible combinations, describing how many jellybeans are pulled out and why, but the method leads to the wrong answer. • Is able to correctly vocalize the logic of the problem.		• Provides the correct answer (If you take out four, at least two have to be the same color.) • Makes very few errors