MULTIPLE RATERS AND HYBRID CONSENSUS IN EVALUATING PRIOR EXPERIENCE FOR A FINANCIAL ANALYST POSITION: ARE RELIABILITY AND EMPIRICAL VALIDITY IMPACTED?

Baugher, Dan
Pace University

Weisbord, Ellen
Pace University

ABSTRACT
The study considers 166 candidates in a State Agency seeking in-line promotions to a financial analyst (FA) position. Subject matter experts (SMEs) rated prior job relevant experience using a training and experience exam (T&E). Supervisors rated current job performance using a behaviorally based performance appraisal (BPAP). The incumbent and promotional positions differed only in level of responsibility and complexity; the behaviors required for success in both positions were identical. Thus, it was possible to use the correlation between the T&E and the BPAP to estimate the empirical validity of the T&E. Three rating approaches are compared: one rater, two raters, and two raters with hybrid consensus. An r of .26 and .21 was found between the T&E and BPAP for the dual rater and single rater conditions, respectively. The difference between these dependent correlations was assessed using a one tail test statistic and found to be significant (Z = 1.68, p < .05). An r of .231 and .116 was found between the T&E and BPAP for the dual rater and single rater conditions, respectively, for those going to consensus (n = 76). This difference was also found to be significant (Z = 1.75, p < .05). Consensus did not improve the T&E by BPAP correlation (p > .05). About 46% of the total scores required a consensus meeting. Interrater reliability improved significantly after consensus, moving from .73 to .90 (p < .05). For scores unified by consensus, it moved from .53 to .89 (p < .05).

INTRODUCTION
Prior training and experience are frequently used in the selection of job applicants in the public sector. Typically, subject matter experts (SMEs) formally assess this information using training and experience exams (T&Es). T&Es are structured application blanks in which candidates describe how their prior training and experience fits the knowledge, skill, and ability (KSA) requirements of the position for which they are applying.

Evaluation of the quality of prior work experience has long been considered a useful method for predicting job performance (Griffin, 1989; Tenopyr & Oeltjen, 1982). Such evaluation has taken many forms over the years, ranging from scored application blanks to structured interviews (McDaniel et al, 1994; Ash & Levine, 1985). In the private sector, employment interviews are the...
most frequently chosen selection instrument for filling open positions (Posthuma et al., 2002) and a review of resumes is common practice for employment decisions at the entry level (Cole et al., 2003). All involve assessment of prior experience. The T&E is a variation on these techniques, evaluating the quality and relevance of prior experience as it relates to KSA requirements of the promotional position.

The quality and relevance of prior experience has been found to relate to future job performance (Sneed et al., 1987; Pulakos & Schmitt, 1995). Some argue on the basis of the “consistency principal” that the best predictor of future behavior is past behavior (Owens, 1976; Wernimont & Campbell, 1968). A meta-analysis of published and unpublished research studies on methods assessing the quality of prior experience (McDaniel et al., 1988) found that the KSA and behavioral consistency methods have a useful degree of predictive validity. The KSA method allows for the evaluation of KSA-based experience relevant to a promotional position through self-ratings, expert ratings, or both. Research also supports the efficacy of scored biographical information inventories in predicting future performance (Cole et al., 2007). While different in format from the current study’s T&E approach, biodata research has revealed a relationship between experience and various measures of performance in keeping with the notion that prior experience matters. T&Es for complex positions frequently use the KSA approach and that is the case for the T&E exams considered in this study.

Given their common use and potential validity for predicting job performance, it is useful to examine how different rating approaches for evaluating prior experience may differ. It takes time and effort to evaluate prior experience, especially when answers are in essay format as they often are for management and complex job positions. The reliability resulting from alternative approaches also sets an upper limit to the empirical validity for such evaluations (Conway, et al., 1995). For these reasons, comparisons of alternative methods for rating prior experience can yield important information for judging their cost benefit.

REVIEW OF THE LITERATURE
A number of studies have found expert ratings to have a positive correlation with job performance. McKillip and Cox (1998) found a positive relationship between expert ratings of job performance and professional certification. Hagman (1998) found that training experts were capable of predicting the rifle marksmanship performance of 51 U.S. Army National Guard soldiers at the marksman, sharpshooter, and expert levels typically used to classify performance. Dipboye (2001) found validity, though weak, between the results of unstructured panel interviews by experts for the training success and job performance of 513 correction officers. These expert ratings by the panels also provided a small incremental value over two paper credentials for predicting officer training success and job performance.

While employees frequently review multiple ratings for performance appraisals and other rating processes used for employment selection, it is typically subject matter experts (SMEs) who rate KSA-based T&Es. The number of raters and the method for aggregating multiple ratings vary. There are single rater systems and multiple rater systems that range from two to as many as ten raters. Methods for aggregating multiple ratings include mechanical methods such as averaging and summing; consensus ratings, which are reached by asking individual raters to work out their differences and agree on a single response; and a hybrid of the two, which combines mathematical and consensual ways to achieve a single response.

The benefits of multiple rater performance appraisals over the more traditional single rater systems have been well researched. Wanguri (1995) found that multiple rater appraisals improved rating accuracy and perceptions of fairness in a meta-analysis of 113 empirical studies on
Multiple ratings can improve the accuracy of prior experience evaluations by making observed scores subject to less random error, so that evaluations of KSA-based experience reach the commonly accepted interrater reliability threshold of .80. In a meta-analysis of job analysis interrater reliability data, Voskuil and Slieker (2002) found that the number of raters needed to reach the .80 reliability standard varied as a function of the content evaluated. When jobs were evaluated for behaviors required for success, a single rater could reach an estimated reliability of .84 or greater. For analysis of KSAs required for success, interrater reliabilities of .80 were more difficult to achieve, needing between five and nine raters to reach that standard.

In their meta-analysis of selection interviews, Conway, et al. (1995) found similar results showing the benefit of multiple raters. Panel interviews yielded reliabilities of .77, on average, in comparison to an average reliability of .53 for separate interviews. Interview structure and number of ratings for the interviews had the greatest impact on interviewer reliability. The researchers also found that subjective combining of multiple ratings did not result in an increase in interrater reliability over that resulting from mechanical combinations of ratings such as averaging or summing.

Finally, while the assessment of a job is different from the assessment of an individual’s prior experience, the research regarding KSA-based job analyses suggests that the use of one rater in the evaluation of an individual's KSA-based experience may not reach the threshold value of .80 advocated by some for strong agreement (Brown & Hauenstein, 2005; Wagner, et al., 2010). The need for multiple raters to reach this threshold value is further suggested by the higher reliabilities found for panel interviews and multiple ratings in interview settings (Conway, et al., 1995).

The research on methods for consolidating the responses of multiple raters is less clear than that on the benefits of multiple raters. Gigone and Hastie (1997) found that the unweighted averages of individual judgments of an organization's properties outperform the consensual approach. For assessment centers, Cohen (1978) argued in favor of consensus as a central part of the rating process. In contrast, Sackett and Wilson (1982) suggested that consensus might not be necessary. For 18 ratings made on 719 individuals, they found that in the absence of consensus the use of a mechanical decision rule could predict consensus results and overall assessment center results with 75.0% and 94.5% accuracy, respectively. As with multiple raters, consensus could improve perceptions by employees that a given rating process is carefully administered (McEvoy, 1990).

Kumar et al. (1993) developed the hybrid approach for creating a single response from multiple informant ratings. In their research, two informants, the sales manager and fleet manager of a large rental company, evaluated dealers on 21 performance assessment items that measured such characteristics as competence, compliance, and adaptation. Using a 7-point Likert scale, each pair of informants evaluated one dealer. The 98 responding pairs were converted into single responses for 98 dealers. Raters were required to use consensus when a difference of two or more points existed between the pair of raters. Using that criterion, consensus was needed for approximately 15 percent of the questions. Wagner, et al. (2010) argue that since the hybrid approach uses consensus only for responses that differ by some specified value, it results in less time and effort and thus is the preferred method for combining multiple rater responses into a single measure.

In sum, support for the value of consensus for T&E ratings is equivocal. For interviews, ratings combined subjectively have not been shown to yield greater reliabilities than those combined
mechanically, through simple addition of the ratings (see, for example, Conway, et al., 1995). For assessment centers, consensus has been advocated but empirical research suggests that that the mechanical combination of ratings in that setting may be as useful as consensus (Sackett and Wilson, 1982). In the development of single measures from multiple ratings, hybrid consensus has been advocated as better than full consensus for combining multiple informant ratings but the benefit of that for reliability coefficients is not clear, and there is little empirical evidence on the relative benefits of averaged ratings versus ratings obtained through full or hybrid consensus for inter-rater agreement and validity (Wagner, et al., 2010).

PURPOSE OF THE STUDY
This study focused on contrasting the results from three approaches to rating training and experience as measured by a T&E: (1) a single expert rater, (2) two expert raters with scores combined mechanically by averaging and (3) two expert raters with scores combined through hybrid consensus.

The goal of the study was twofold: to learn the impact of number of raters on the correlation between prior experience and current job performance, and to learn the impact of using average scores versus hybrid consensus on the correlation between prior experience and current job performance. Job performance, measured after the T&E was completed, was considered a proxy for candidate performance in the promotional position because the behaviors needed for success in the promotional position are identical to those required in the incumbent position. Thus, this correlation is suggestive of the potential empirical validity of the T&E for future job performance.

HYPOTHESES
Two hypotheses were tested. In Hypothesis 1, we predicted that the correlation between prior experience and supervisor ratings of current job performance would increase as a function of the rating approach used. Specifically, it was expected that this correlation would improve when moving from one rater to two raters and when moving from the mechanical averaging of scores to hybrid consensus to improve rater reliability.

In Hypothesis 2, we predicted that the intraclass reliability for two raters would improve following a consensus meeting. This was expected for the entire sample even though consensus was required only for total T&E scores with differences greater than a specified amount. We expected reliability to improve because, in using the consensus procedure to review meaningfully divergent ratings, raters must rethink their application of standards, reducing the random error from irregularities in the application of standards as interpreted by different reviewers.

H1: The correlation between T&E expert scores and current job performance ratings will increase as a function of the rating method.
   H1a. The two-rater condition will yield a higher correlation than the one-rater condition.
   H1b. The hybrid consensus condition will yield a higher correlation than the two-rater condition.

H2: The intraclass correlation coefficient for T&E expert scores resulting from hybrid consensus will be of a greater magnitude than the coefficient resulting from the average of two raters without consensus.

SUBJECTS
Subjects were 166 candidates for promotion to a financial analyst (FA) position within a State Agency. Data was obtained from seven promotional examinations over a 12-year period. In order to apply for promotion, candidates had to be in the incumbent position for at least one year.
Both the promotional and incumbent positions required the same KSAs and behaviors for successful performance. The promotional position differed only in its complexity and level of responsibility. This is the case for many in-line promotions, such as those in the trades and professional positions such as professor.

**INSTRUMENTS AND RATING PROCESS**

Two instruments were used in this study. The T&E assessed the quality of prior work experience. The Behavioral Performance Appraisal for Promotion or BPAP provided an indicator of the candidate's current job performance on behaviors identical to those required in the promotional position.

**Evaluating prior experience.** Prior experience was evaluated by asking candidates to describe their highest quality experiences for the specified KSA. Experience could be acquired within or outside the Division, as long as it related to the KSAs. Responses were open-ended and were often from one to three pages in length. No page limit was set, though research not cited here showed that page length did not correlate with the evaluations of prior job-related experience. The T&E was developed through a rigorous methodology to assure that the KSAs were important in the promotional position and candidates could acquire the necessary experience in their current position (Lawshe, C., 1975; New York State Department of Civil Service, 1976).

Over the 12-year time period, the T&E was comprised of either 26 or 22 items with 39 percent of the candidates rated on 26 items and 61 percent rated on 22 items. For ease of comparison and understanding, all scores were normalized to range from 0 to 100 by dividing the rater raw score by the maximum score possible. For the 26-item and 22-item T&Es, the maximum possible was 78 and 66, respectively.

SME raters were in the next grade or higher but not a supervisor of any candidate seeking promotion. All had to be familiar with the incumbent and promotional position. When an SME said he or she was familiar with the candidate, they were given a different T&E to evaluate. Since the T&Es were rated by two SMEs, SMEs had to evaluate 332 T&Es. Forty-five SMEs rated the T&Es in differing pair combinations. They were assigned from 1-11 candidates to evaluate as a first or second rater depending upon their availability and the candidates rated. On average, each rater evaluated 7.4 T&Es and 3.7 candidates.

SMES were provided with a comprehensive training program and a scoring booklet. The scoring booklet was designed by SMEs within the Division to provide examples of responses for each of the four ratings possible for a KSA item. The four rating possibilities are shown below.

<table>
<thead>
<tr>
<th>Score Value</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No relevant training, education, or experience.</td>
</tr>
<tr>
<td>1</td>
<td>Education only, training only, and/or limited job experience.</td>
</tr>
<tr>
<td>2</td>
<td>Typical job experience.</td>
</tr>
<tr>
<td>3</td>
<td>Unusually superior and expert job experiences.</td>
</tr>
</tbody>
</table>

Each SME was blind to the ratings of the other SME. If the total score resulting from the two raters differed by more than seven points (about 10% of the maximum total possible), a consensus...
meeting was required to push for bringing the two scores within seven points of each other. This “rule of seven” resulted in 45.8% of the T&Es requiring a consensus meeting. Meetings typically lasted about 45 minutes. While it was not required that raters bring their scores within 7 points, almost all did so.

The single rater T&E score was based on the rating of one of the two raters. To minimize any potential order effect for the raters, this T&E score is the result of randomly selecting one of the two ratings for each candidate. This selection was made using Bernoulli distribution to select one of the two scores to stand in for the one-rater score. This resulted in 51% and 49% of the rater 1 and rater 2 scores being the source of this score, respectively.

The T&E score for the two-rater condition was the average score assigned by the two raters. The hybrid consensus T&E score was the average score for the two raters following any needed consensus meeting.

Performance appraisal instrument, scores, and implementation. A behavior-based performance appraisal comparable to the behavioral observation scales (BOS) developed by Latham and Wexley (1981) was developed for the candidate position. The appraisal contained 65 percent positive and 35 percent negative behaviors related to performance in the promotional position and observable in the candidate’s current position. The behaviors were identified through rigorous analysis of hundreds of behaviors performed in both the candidate and promotional positions. The performance appraisal ranged in length from 45 to 46 behavioral-based items as the result of reevaluations of the instrument at different points in time.

The basis of the five-point rating system was the percentage of instances that a candidate engaged in a behavior compared to the number of opportunities the candidate had to engage in the behavior. Higher scores indicated that the candidate was performing as expected while lower scores indicated that performance was less than expected. For negative items, a high score meant the candidate avoided negative behaviors. For positive items, a high score meant the candidate engaged in positive behaviors.

Each candidate’s current supervisor evaluated the candidate’s performance. The score was normalized to a 100-point basis for ease of comparison and is referred to here as the BPAP in this study.

EMPIRICAL VALIDITY
In the assessment of H1, current job performance ratings served as a proxy measure for a candidate's performance in the promotional position. The use of current job performance as a proxy in this situation was reasonable because the incumbent position requires the same job behaviors for success as the promotional position. Current job performance ratings were available for all subjects. Unlike studies where few may be promoted, substantially limiting the degree to which a T&E can relate to performance, this study permitted an assessment of how prior experience might relate to future performance for all candidates, irrespective of how they scored on the T&E.

The correlation between the T&E and current job performance is one approach to assessing the empirical validity of the T&E. For H1, the empirical relationship was expected to improve when more than one rater was used to assess prior experience and, once again, when the reliability of two raters was improved through hybrid consensus.
INTER-RATER AGREEMENT AND RATER RELIABILITY

In the assessment of H2, the single measure intraclass correlation was used to assess the reliability of the total score resulting from the two ratings. The intraclass correlation does not require that the raters be equivalent forms (Bartko, 1966). The coefficient used in this study is often referred to as the one-way ANOVA intraclass correlation or ICC(1).

Choice of an intraclass reliability coefficient depends upon how raters are used in a study. In this study, all candidates were rated by two SMEs but the same pair of SMEs did not rate each candidate. Thus, the one-way ANOVA approach or ICC(1) was appropriate as raters cannot be a factor (Bartko, 1976; Bartko, 1978; Shrout and Fleiss, 1979). ICC(1) can be interpreted as indicating the level of intraobserver consistency one can expect in the future, should the same background be evaluated using the same coding scheme with observers of equivalent training.

In the assessment of H2, the intraclass reliability coefficient for average ratings was also used in testing the hypothesis. This is sometimes referred to as the Spearman-Brown prediction or ICC(2). ICC(2) assesses the reliability of average ratings rather than the reliability of a single rating. If another random sample of raters were to rate the same candidates, ICC(2) provides the correlation between averaged ratings that could be expected from the two sets of raters (Bartko, 1976; Winer, 1971).

The intraclass reliability coefficients compared in H2 were correlated correlation coefficients because the same subjects were rated in each condition. This complicated the evaluation of the statistical significance of their difference. Donner and Zou (2002) compared several approaches for testing the equality of dependent intraclass correlation coefficients, including Fisher’s Z test, the Konishi-Gupta modified Z test, the likelihood ratio test and the Alsawalmeh-Feldt F test using Monte Carlo simulation studies. Unfortunately, these tests are not standard, not easy to perform, and their power needs further investigation.

As a result, H2 was evaluated by comparing the overlap in the 95% confidence intervals for the coefficients before and after consensus. In this approach, the absence of any overlap is a strong indicator that the coefficients differ more than chance would predict (Lu and Shara, 2007; Payton, et al., 2003). Some argue that this approach is conservative when 95% confidence intervals are compared. Payton, et al. (2003) suggest that when the standard errors are approximately equal, use of an 84% interval size to check for overlap better approximates an alpha of .05. In evaluating H2, the more conservative 95% intervals were compared to test the hypothesis that the intraclass reliability coefficients for the study increased more than chance would predict following consensus.

RESULTS

Table 1 provides descriptive statistics for the T&E and BPAP across the three rating conditions for all subjects, those going to consensus, and those not going to consensus. About 46% of the candidate scores required a consensus meeting. An analysis of the mean and variance differences shows that there are no significant differences between the scores of subjects going to a T&E consensus and those not requiring consensus ($\rho > .05$).

H1a predicted that the correlation between evaluations of prior experience through a T&E and ratings of job performance would improve when two experts were used to derive the T&E score. H1b predicted that the same would hold true when hybrid consensus was used to derive the score. This study tested the significance of the difference between these correlations using the Fisher's Z-transformation following the method suggested by Meng, Rosenthal, and Rubin (1992) for dependent correlations.
Table 1
Mean, Standard Error, and Standard Deviation for T&E and BPAP Across the Three T&E Rating Approaches

<table>
<thead>
<tr>
<th>Sample</th>
<th>T&amp;E/BPAP Correlation</th>
<th>T&amp;E/BPAP Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>N = 166</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>S.E.</td>
</tr>
<tr>
<td></td>
<td>61.6</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>17.0</td>
</tr>
<tr>
<td>Consensus</td>
<td>Mean</td>
<td>60.1</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>16.4</td>
</tr>
<tr>
<td>No Consensus</td>
<td>Mean</td>
<td>62.9</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>17.4</td>
</tr>
<tr>
<td>BPAP</td>
<td></td>
<td>89.2</td>
</tr>
</tbody>
</table>

Table 2 provides the correlation between the T&E and BPAP in the three conditions for all subjects and for those subjects where consensus was required. For all subjects (n = 166), the correlation between the T&E and BPAP was greater than would be expected due to chance at the .05 level, using a two-tail test for all three T&E scoring conditions. For subject scores going to consensus (n = 76), the correlation between the T&E and BPAP was greater than what might be expected due to chance only in the two-rater condition and not for the one-rater condition or consensus condition.

The differences between the T&E/BPAP correlations shown in Table 2 were tested using a one-tail statistic to increase power with the understanding that it might foster a greater chance of a false positive in assessing the differences. The correlation between the T&E and BPAP for the two rater condition was significantly greater than that found in the one rater condition for all subjects (Z = 1.68, p < .05) and for consensus only (Z = 1.75, p < .05). The correlation between the T&E and BPAP for the two rater condition was not significantly different than that found for the consensus condition for all subjects (Z = -.77, p > .05) or for consensus only (Z = -.85, p > .05). It should also be noted that the correlation between the one rater T&E score and two rater score was .923 for all subjects and .854 for consensus only subjects.

Table 2
Intercorrelation Between T&E Score and BPAP by T&E Rating Condition for All Subjects and Consensus Only Subjects

<table>
<thead>
<tr>
<th>T&amp;E Rating Approach</th>
<th>T&amp;E/BPAP Correlation All Subjects (n = 166)</th>
<th>T&amp;E/BPAP Correlation Consensus Only (n = 76)</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Rater</td>
<td>.210**</td>
<td>.112</td>
</tr>
<tr>
<td>Two Raters</td>
<td>.260***</td>
<td>.231*</td>
</tr>
<tr>
<td>Hybrid Consensus</td>
<td>.251***</td>
<td>.198</td>
</tr>
</tbody>
</table>

* p < .05, ** p < .01, *** p < .001
Thus, H1 is partially supported. The two-rater condition (H1a) was found to improve the empirical relationship between the T&E and BPAP compared to using one rater to evaluate prior experience. To be sure, the correlations are not large but it is important that for the consensus condition the use of two raters brought the relationship between the T&E and BPAP into statistical significance, suggesting that it is not due to chance. Contrary to findings for H1a, the consensus condition (H1b) was not found to improve the relationship between the T&E and BPAP compared to the two-rater condition without consensus.

H2 predicted that interrater reliability for the T&E scores would benefit significantly from the hybrid consensus meetings. The ICC(1) intraclass coefficient was .73 prior to consensus and .90 following consensus. The 95% confidence interval for the pre-consensus coefficient is .651 to .794 and the 95% interval for it following consensus is .865 to .924. There is no overlap between the two intervals showing that the difference uncovered is more than chance would predict (p < .05). As neither interval includes a 0, they are statistically significant correlations (p < .05). This is a conservative test for an alpha of .05, since the standard error for two raters before and after consensus is almost identical, as shown in Table 1.

The ICC(2) intraclass reliability coefficient was also computed before and after consensus. This provides an estimate of the interrater reliability that can be expected for the average scores generated by a pair of raters. ICC(2) for the FA position before consensus was .84 and .95 following consensus. As was the case for the ICC(1) intraclass reliability coefficient, the 95% confidence intervals for these coefficients do not overlap. They are .789 to .885 before consensus and .928 to .961 following consensus suggesting a significant difference (p ≤ .05) at a conservative level, given that the standard error before and after consensus was so close. As neither interval includes a 0, they are statistically significant reliability coefficients (p ≤ .05).

While not a part of H2, the improvements shown for ICC(1) for the subset of scores going to hybrid consensus were also compared. About half the T&Es in the FM and UM positions were sent to a consensus meeting. For these T&E scores the intraclass coefficient of reliability was .53 before consensus and .89 after consensus with 95% confidence intervals of .344 to .671 before consensus and .832 to .929 after consensus. The absence of overlaps between these confidence intervals indicates that the correlations are significantly different (p ≤ .05). The absence of a 0 in the intervals shows that the reliability coefficients are statistically significant (p ≤ .05).

Thus, H2 is fully supported. The implementation of hybrid consensus contributed significantly to improvements in reliability for the overall system and, as would be expected, for those who had total scores that were not considered sufficiently close. This increase in reliability through consensus did not significantly improve the correlations between the T&E and the BPAP found prior to consensus despite the increase in reliability resulting from consensus.

**DISCUSSION**

This study focused on the complex rating process that takes place when prior experience is evaluated against KSA-based standards for success in a future position and when the information available to the rater is in essay format. Evaluations were made using a T&E examination, which is a common practice in State government. Two hypotheses were suggested.

First, the study investigated whether improvement in the empirical relationship between ratings of prior job-related experience and job performance occur as a result of using two raters instead of one, and also whether using consensus to get a single score for two raters improves reliability beyond that using mechanical averaging to get that single score.
While it was not possible to track future performance for the candidates promoted, it was possible to assess job performance for all candidates in their current position for behaviors that were identical to those required in the future position. The authors have in this paper argued that this provides a viable measure of the empirical relationship of ratings of prior job-related experience with future performance. Due to the high degree of similarity between the incumbent and promotional positions as well as the rigorous assessment of performance compared to ordinary annual evaluations, having this data provided a rare opportunity to make such an assessment.

Second, this study was interested in how the interrater reliability for all candidates changed as a result of hybrid consensus, a method used when some total scores are close enough to use the mechanical averaging method of reaching a single score, and other total scores are sufficiently diverse to require using a consensus approach to reach a single score. The authors expected that hybrid consensus would improve reliability for the entire system. The focus was on the whole system because assessment of the entire group, not subsets of candidates, is the key concern of assessment systems. The finding that that the overall system improved was not surprising, since consensus improved reliability and 46% of all candidate scores required a consensus meeting.

The correlation between the T&E measure of prior experience quality and job performance when using two raters instead of one was found to provide a small amount of improvement. Though not as large a correlation as expected, the finding is useful. Content validation is typically used for T&E exams, and the higher cost in logistics and experts’ time when using two raters instead of one is of consequence. The statistically significant increase found in the two-rater condition supports the potential empirical validity of a T&E and provides a rationale for the higher costs of the two-rater methodology.

The intraclass reliability coefficients show that SME ratings of prior experience resulted in moderate agreement among the raters without using consensus meetings. The coefficient of .73 suggests that raters can make moderately consistent ratings of KSA-related prior experience from essay answers. With hybrid consensus, interrater reliability for the process increased considerably, moving from moderate levels of agreement to high levels of agreement, yielding an intraclass correlation of .90. This level of reliability required approximately 45 additional minutes of rater effort.

It is interesting that this degree of improvement in reliability did not bring about concomitant improvement in the empirical relationship between the T&E and job performance ratings. While not significantly different from the non-consensus dual rating correlations, the T&E by BPAP correlations moved in the direction of getting smaller with the use of consensus. One possible reason may be that the focus on total score consistency and not item consistency in the hybrid consensus condition provided a “quick fix.” On the other hand, the result could simply be chance variability in a situation where the performance variable ratings were considerably restricted in range relative to the ratings of prior job-related experience. Performance ratings are well known for their restriction in range.

The results of this study, both anticipated and not anticipated, can be judged as heartening for the use of complex evaluation methods. One encouraging outcome is evidence of the remarkable ability of busy managers to evaluate complex write-ups of prior experience with at least a moderate degree of agreement. Candidate submissions ranged from 30 to as many as 100 pages in length and the KSAs against which that experience had to be evaluated were complex. For those who must use a single rater to make such evaluations, the study suggests that ratings of this complex material is at least moderately high in interrater reliability at 73%. Hybrid consensus,
where consensus was used for total scores that differed by 10% or more, increased interrater reliability substantially to 90%.

Another reassuring result is the finding of empirical validity for expert ratings of prior experience and performance whether one rater or two raters was used, though the latter modestly improved the relationship. T&E scores are rarely related to performance of any sort as their validation typically focuses on content validity and not empirical validity. Indeed, it is uncommon for field research to find empirical validity between assessments of prior experience and important criteria, such as future performance, whether the assessment tool is an interview, an application blank review, or a T&E. The current study found a correlation between prior experience and performance to exist in a situation where an extremely complicated rating process was used to evaluate essay write-ups of experience, and supervisor ratings were used to assess performance in an on-the-job setting and not a manufactured one. The results suggest that some relationship between quality experience and performance can be expected in future research.

REFERENCES


