Proceedings Submission

Hawaii International Conference on Business: May 23-26, 2013

Title: Using Multiple Raters and Consensus to Evaluate Prior Experience for Middle Managers: Do These Approaches Improve the Association of Experience with Job Performance?

Paper Author: Dan Baugher

Author Names/Affiliations: Pace University

Author Address: Dan Baugher

150 Nassau Street, APT 11D

New York, N.Y. 10038

Email: [dmbaugher@aol.com](mailto:dmbaugher@aol.com)

Full Paper Follows:

**Using Multiple Raters and Consensus to Evaluate Prior Experience for Middle Managers: Do These Approaches Improve the Association of Experience with Job Performance?**

**Dan Baugher, Pace University**

**ABSTRACT**

This study considers whether the use of multiple raters and consensus in the evaluation of prior experience improves the relationship between prior experience and current job performance for a group of middle managers. From 1983 through 2012, a State Agency administered the same system for promoting middle managers to top management. Candidates completed a training and experience exam (T&E) where they described how their experience related to the KSA requirements for the promotional position and their current job performance was rated by their supervisor three months after they completed the T&E, with an appraisal developed specifically for the exam process. The incumbent and promotional positions differed only in level of responsibility. Thus, the KSAs assessed by the T&E were important for success in the incumbent position, making it possible to assess the empirical validity of prior experience for predicting current performance. The system was administered on eight occasions for 242 candidates. Individual rater and pre-consensus data was retained on four occasions for 127 candidates: 1983, 1998, 2007, and 2012. Three rating approaches are compared: one rater, two raters, and two raters with hybrid consensus, where a consensus meeting took place only if the total score assigned by the raters differed by a fixed amount. The single rater and dual rater correlations of experience with job performance were .412 and .504, respectively, across all subjects and .415 and .575, respectively, for those going to consensus (n=46). All four correlations are statistically significant (*p* < .001). Using a one-tail test, the difference between these dependent correlations for both samples was also found to be statistically significant (*p* < .01). The correlation with performance for hybrid consensus was not significantly different (*p* > .05) from that found for two raters without consensus. The results suggest that the use of two raters can improve the empirical validity of a T&E exam for predicting job performance but additional improvement from consensus is not likely.

**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 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 assess KSA-based experience using Subject Matter Experts (SMEs) to make the evaluations and that is the case for the T&E exams considered in this study. This differs from task-based T&Es which require that job candidates only check off tasks that they have engaged in.

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 impact the empirical validity of experience for predicting job performance. It takes time and effort to evaluate the prior experience of managers. This information is invariably provided in an essay format and can involve many pages. 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.

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 process, which combines mathematical and consensual ways to achieve a single response from more than one rater.

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 performance appraisals. Latham and Wexley (1982) argued that multiple raters minimized the weakness of individual ratings. Bernardin and Beatty (1984) found that multiple raters provided an improved legal defensibility over ratings from one person.

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, Voskuijl and Sliedregt (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.

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). Apart from reliability, there is the issue of empirical validity. While the evidence suggests that multiple raters may improve validity and consensus may have no impact, there is, in fact, little field research to substantiate or counter either possibility.

**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 following hybrid consensus. The term “hybrid consensus” reflects the fact that all scores did not go to consensus. Rather, rater scores were subject to consensus meetings only when the raters’ total scores differed by more than a fixed amount.

The goal of the study was twofold: (1) to determine the impact of number of raters on the relationship of expert ratings of prior job experience with current job performance and (2) to determine the impact on this relationship of having two raters’ scores subject to hybrid consensus compared to that of two raters’ scores with no consensus meeting. In this study, empirical validity of the T&E was deemed to exist when a correlation between expert scores of prior experience and supervisory ratings of current performance was uncovered.

**HYPOTHESES**

Two hypotheses were tested. Hypothesis 1 predicted that the correlation between prior experience and supervisor ratings of current job performance would improve through the use of two raters to evaluate prior experience over that found for a single rater. Hypothesis 2 predicted that the correlation between prior experience and supervisor ratings of current job performance would improve through the use consensus for scores differing by a set amount (referred to as hybrid consensus) compared to that found for two raters without any consensus. That is, the trajectory of improvement was expected to improve first as a result of moving to two raters from one rater and again by moving from two raters to two raters with hybrid consensus.

**H1:** The correlation between expert ratings of prior experience quality and current job performance ratings will be larger for expert scores derived from two raters than for scores derived from a single rater.

**H2:** The correlation between expert ratings of prior experience quality and current job performance ratings will be larger when the score from two expert raters is determined after a hybrid consensus meeting than when the score is determined without a consensus meeting.

**SUBJECTS**

Subjects were 127 candidates for promotion to a top management position within a State Agency. All candidates were middle managers seeking an in-line promotion where the focus of their efforts and job context in the promotional position would be the same as in their current position but with added responsibility. All managers were responsible for, managing budget analysts, implementing Executive Office directives and policies for the State budgets, negotiations with agencies on budget implementation, and assuring that budget studies were complete, accurate, and timely. In the promotional position, they would continue these activities but with oversight of larger budgets, more complex State agencies, and move from management of budget analysts to management of those in their own middle-management positions.

From 1983 through 2012, the Agency administered the same system for promoting these middle managers to top management with minimal change. The process was implemented on eight occasions for 242 candidates. Individual rater and pre-consensus data was retained for 127 candidates across four administrations: 1983, 1998, 2007, and 2012 with a sample size of 33, 23, 26, and 45, respectively. 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. This meant that the KSAs necessary for successful performance in the promotional positions were the same KSAs required for successful performance in the incumbent position. This was a result of the in-line nature of the promotion; candidates would continue to manage the same processes and engage in many of the same activities upon promotion, but with the added responsibility typical of a promotion to top management. Likewise the behaviors required for success in the incumbent and promotional position were the same.

**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 (BPAP) provided an indicator of the candidate's current job performance on behaviors identical to those required in the promotional position.

**EVALUATING PRIOR EXPERIENCE.** To evaluate prior experience, candidates were asked 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 sited 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).

Over the close to 30-year time period, the T&E was comprised of either 19 or 13 items with a reduction in items taking place in 2001. Thus, 44% of the candidates in this study were rated on 19 items and 56% were rated on 13 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 19-item and 13-item T&Es, the maximum possible was 57 and 39, respectively.

SME raters were in the next grade or higher but not a supervisor of any candidate seeking promotion. All were high-level, senior managers in the Agency and all had to be familiar with the incumbent and promotional position. In later years, almost all SMEs had gone through the examination process to reach their current top management position and were, as a result, very familiar with the system. 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 254 T&Es. Thirty-one SMEs rated the T&Es in differing pair combinations. They were assigned from 1-8 different candidates to evaluate as a first or second rater depending upon their availability and the candidates rated. On average, each rater evaluated 8.2 T&Es and 4.1 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.

Score Value Meaning

0 No relevant training, education, or experience.

1 Education only, training only, and/or limited job experience.

2 Typical job experience.

3 Unusually superior and expert job experiences.

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. A consensus meeting was required to push for bringing the two scores within seven points of each other. This “rule of seven” resulted in 36.4% 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. Since the Agency decided to continue using this same standard when the T&E shrunk to 13 items in order to reduce the number of consensus meetings that top management had to attend, the percentage of consensus meetings differs for the two administrations held before 2001 from those administered after that date. For those administered prior to 2001, 41.4% of the T&Es required a consensus meeting. For those administered after this date, 32.4% of the T&Es required consensus.

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 for each exam administration.

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. As noted, consensus was not needed for all T&Es.

**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 33 behaviorally-based items related to performance in the incumbent and promotional position. There were 33.3% and 66.7 % negative and positive behaviors, respectively. All were 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 appraisal was referred to as the Behavioral Performance Appraisal for Promotion (BPAP) and retained the same number of items for the close to 30-year period over which the exams were administered, with only minor changes to the items themselves.

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.

**INTER-RATER AGREEMENT AND RATER RELIABILITY.** The reliability of the total score resulting from the two ratings was assessed by the single measure intraclass correlation. 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.

The intraclass reliability coefficient for average ratings was also used as a measure of reliability. 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 obtained before and after consensus were correlated correlation coefficients because the same subjects were rated before consensus and after consensus. 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, the statistical significance of the difference in the reliability coefficients 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. Nonetheless, it was decided that the more conservative 95% intervals would be compared to check if statistically significance improvements in reliability resulted from the hybrid consensus approach.

**EMPIRICAL VALIDITY.** Both H1 and H2 focus on the empirical validity of expert ratings of prior experience for predicting current job performance. The approach to assessing empirical validity is that of concurrent validity since both measures were taken around the same time though the T&E was completed three months before performance was evaluated. It should be noted that almost all experiences described in the T&E took place prior to the performance rating period with some taking place a year or, often, several years and longer in the past. This was possible because candidates could list any prior experience that could be verified in their past. The T&E was not limited to what they had done while in their current position though all KSAs could be achieved in their current position.

The correlation between the T&E and BPAP is used to assess the empirical validity of the T&E for predicting current job performance. For H1 and H2, this relationship was expected to improve when more than one rater was used to assess prior experience and, once again, when two raters went on to consensus as needed (referred to as hybrid consensus).

**RESULTS**

Table 1 provides descriptive statistics for the T&E and BPAP across the three rating conditions for all subjects. As noted earlier, about 36% of the candidate scores required a consensus meeting. A comparison of the two-rater T&E scores for those requiring a consensus meeting and those who did not require a meeting showed no significant difference between the means and variances of the two groups (*p* > .05). There was also no significant difference (*p* > .05) in the means and variances of the BPAP scores for the two groups.

Interrater reliability for the T&E scores before and after hybrid consensus was compared. The ICC(1) intraclass coefficient was .676 prior to consensus and .872 following consensus. The 95% confidence interval for the pre-consensus coefficient is .570 to .760 and the 95% interval for it following consensus is .824 to .908. 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 (See 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. The ICC(2) intraclass coefficient before consensus was .807 and .932 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 .726 to .863 before consensus and .903 to .952 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 also statistically significant reliability coefficients (*p* < .05).

**Table 1**

**Mean, Standard Error, and Standard Deviation for T&E and BPAP Across**

**One-Rater, Two-Rater and Hybrid Consensus T&E Rating Approaches**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sample |  | One-Rater T&E | Two-Rater T&E | Consensus T&E | BPAP |
| All | Mean | 63.4 | 64.6 | 64.6 | 92.0 |
| N = 127 | S.E. | 1.6 | 1.5 | 1.5 | .7 |
|  | S.D. | 18.2 | 17.1 | 17.3 | 8.1 |
| Consensus | Mean | 62.2 | 64.2 | 64.1 | 91.5 |
| N = 46 | S.E. | 3.0 | 2.7 | 2.8 | 1.2 |
|  | S.D. | 20.9 | 18.5 | 19.0 | 8.4 |
| No Consensus | Mean | 64.2 | 64.8 | 64.8 | 92.4 |
| N = 81 | S.E. | 1.8 | 1.8 | 1.8 | .9 |
|  | S.D. | 16.5 | 16.3 | 16.3 | 7.9 |

The impact of consensus on the reliability of those scores requiring consensus was also evaluated using ICC(1). About 1/3 of the T&Es were sent to a consensus meeting. For these T&E scores, the ICC(1) intraclass coefficient of reliability was .464 before consensus and .837 after consensus with 95% confidence intervals of .209 to .660 before consensus and .726 to .905 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). Similar improvements were uncovered for the ICC(2) coefficient which moved from .634 to .911 (*p* < .05).

Thus, consensus had the intended impact on interrater reliability. It improved the reliability of the system, moving the reliability for T&E scores from .676 to .872. For T&E scores requiring consensus, the reliability coefficient moved from .464 to .837. In sum, 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.

H1 predicted that the correlation between evaluations of prior experience and ratings of job performance would improve when the average of two expert T&E scores was compared to one expert T&E score. H2 predicted that this correlation would further improve when the T&E score of two experts using hybrid consensus was compared to the average T&E score of two experts without hybrid consensus. 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 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 = 127), the correlation between the T&E and BPAP was greater than would be expected due to chance at the .001 level, using a two-tail test for all three T&E scoring conditions, ranging from .412 to .510. For subject scores requiring consensus (n = 46), the correlation between the T&E and BPAP was also greater than what might be expected due to chance at the .001 level for the two rater and hybrid consensus condition and at the .01 level for the single rater condition ranging from .415 to .590.

**Table 2**

**Intercorrelation Between T&E Score and BPAP by T&E Rating Condition**

**for All Subjects and Consensus Only Subjects**

|  |  |  |
| --- | --- | --- |
| T&E Rating Approach | T&E/BPAP Correlation  All Subjects  (n = 127) | T&E/BPAP Correlation Consensus Subjects  (n = 46) |
| One Rater | .412\*\*\* | .415\*\* |
| Two Raters | .504\*\*\* | .575\*\*\* |
| Hybrid Consensus | .510\*\*\* | .590\*\*\* |

\* *p* < .05, \*\* *p* < .01, \*\*\* *p* < .001

The differences between the T&E/BPAP correlations shown in Table 2 were tested using a one-tail test 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 = 2.50, *p* < .001) and for consensus only (Z = 1.96, *p* < .05). The correlation between the T&E and BPAP for the hybrid consensus condition was not significantly different from that found for the two rater condition, without consensus, across all subjects (Z = .78, *p* > .05) or for T&Es requiring consensus (Z = .78, *p* > .05, n=46)). It should also be noted that the correlation between the one rater T&E score and two rater score was .892 for all subjects and .802 for ratings requiring a consensus. While the improvements consensus might bring for scores going to consensus impact the overall system, it is the empirical validity of the entire system that is often the primary concern and a key reason why impact of consensus on the overall system are presented.

H1 was supported. The two-rater condition improved the empirical relationship between prior training and experience ratings (the T&E) and current job performance (the BPAP) moving the correlation between the two from .412 to .504. Where consensus was required (n=46), two raters also improved the correlation moving it from .415 to .575.

H2 was not supported. Consensus did not improve the relationship between the T&E and BPAP even though it is clear that the reliability of the T&E improved through the use of consensus. Across all scores, the correlation between the T&E and BPAP was .504 for two raters and .510 for two raters with hybrid consensus. Where consensus was required, the correlation between the T&E and BPAP was .575 for two raters and .590 for two raters with 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 the empirical relationship between ratings of prior job-related experience and job performance would improve with the use of two raters instead of one rater. Second, it assessed whether using consensus to get a single score for two raters for scores apart by a fixed amount would improve this empirical relationship beyond that provided through mechanical averaging to get that single score, without a consensus meeting to bring more divergent scores closer together. This process was referred to as hybrid consensus because all scores did not go to consensus.

While it was not possible to track future performance for any candidate promoted, current job performance for all candidates taking the exam was available. Further, the behaviors required for success in the incumbent position were also the behaviors required for success in the promotional position and performance was assessed using this special performance appraisal system and not the far simpler, annual appraisal system which was tied to merit pay. Since the KSAs required in the promotional position and included on the T&E were also identical to the KSAs required for success in the incumbent position, it was possible to determine if these KSA-based experiences related to current performance, thereby providing an estimate of the concurrent validity of prior experience for these middle management positions.

The study found empirical validity for expert ratings of prior experience and current performance for the middle management positions under consideration. All three rating approaches maintained moderate empirical validities but only the move to two raters resulted in a significant and substantive improvement in the correlations. T&E scores are rarely related to performance of any sort as their validation typically focuses on content validity and not empirical validity. So, the results add information in an area where little is actually known. Indeed, it is uncommon for field research to find empirical validity between assessments of prior experience and current or future performance, whether the assessment tool is an interview, an application blank, or a T&E. The increase in empirical validity for the use of two raters over one rater provides a clear rationale for the more complex logistics and use of expert time required for two raters instead of one.

The results also suggest that there is benefit to the use of complex evaluation methods. The study demonstrates the ability of top managers to evaluate complex descriptions of prior experience with at least a moderate degree of agreement. Descriptions of prior experience ranged from 30 to 100 pages in length and were evaluated against complex KSAs. Study results suggest that even a single rater’s evaluations can be expected to have a moderately high degree of reliability

The use of consensus meetings clearly improved interrater reliability, moving the intraclass correlation of .807 to .932 for the entire sample. The improvement for scores requiring consensus was greater, moving the intraclass correlation from .464 to .837 as might be expected; it was essentially impossible for interrater reliability to fail to improve from this process and a main reason for its implementation. This improvement in reliability required approximately 45 additional minutes of rater effort. However, the improvement in reliability did not bring about a concomitant improvement in the empirical relationship between the T&E and job performance ratings across the entire system or for those scores going to consensus. Consensus simply had no impact on empirical validity.

One reason that consensus may not have improved validity might be an excessive emphasis on total score consistency. This requirement may have resulted in each rater focusing on scores where they gave the item a + or – in their notes but not necessarily any focus on scores that diverged by a great deal where issues of accuracy and validity may have been more important. Indeed, no specific instructions were provided other than to reach an agreement that allowed scores to meet the standard. As a result, score differences were reduced where there was gray, thus improving reliability, but not necessarily where substantive differences in opinion may have existed. This was not formally assessed but was observed by the first author over the time period.

It should be emphasized that individual rater ratings, when two or more raters are used, are not typically available for any length of time after such processes are concluded. Whether they are retained at all, depends on the system in place. Pre-consensus ratings are less often available (if not wholly unavailable) since it is quite common for ratings to be changed without any record of what was changed or by whom. What matters in such situations is the final score produced by the raters. The first investigator created and oversaw this system since its inception in 1983 and was granted access to this very rare data. It was fortuitous that the information was retained at widely space periods of time in the close to 30 years that this process was administered. It should be noted that these middle managers were key figures in the implementation of the State’s budget which is an especially important concern in this era.

Also, the process changed in no significant way during this long time period which is also uncommon, some might say unheard of, providing a rare opportunity to examine field data on the relationship of prior experience to current job performance where no changes in instrumentation or process took place. It is in a way a testament to the success of the system for top management and their view of its validity, as demonstrated in their view of the performance of those promoted over many administrations and Agency Directors. It was a rare situation where the process became part of the culture and was maintained by each forthcoming generation of top managers as they also had successfully gone through the process.

Those who use T&Es or other measures of prior experience can take heart in knowing that, in this setting over decades, that prior experience related to job performance on tasks related to the KSAs assessed with the T&E even though many different supervisors, many different SMEs, and substantially different budget environments faced these middle managers in their work. Also, those faced with a decision on whether to use one rater or two raters to evaluate the essays often generated by such a process or similar processes, should lean toward using two raters, if at all possible, since empirical validity is likely to improve.

On the other hand, those contemplating the use of consensus to gain even greater agreement between raters may want to proceed with more caution. While this process seemed like it would result in better validity when this system was implemented in 1983, it did not result in the expected improvement though it did give greater comfort to candidates though, truth be told, the original concern was as much about assuring interrater reliability as validity.

In hindsight, the problem with consensus may have been the way it was implemented and, at least in this case, change might have been beneficial. Perhaps, if raters had been forced to consider items where they held strong divergent views to reduce the disparity between their total scores, an improvement in empirical validity might have been uncovered and reliability would still have improved. Unfortunately, that was not a requirement in the current system and the assessment of the benefit of such an approach will have to take place in another setting at another time.

**REFERENCES**

Ash, R., and E. Levine (1985). “Job Applicant Training and Work Experience Evaluation: An Empirical Comparison of Four Methods.” *Journal of Applied Psychology,* 70(3), 572-576.

Bartko, J. J. (1966). “Intraclass Correlation Coefficient as a Measure of Reliability.” *Psychological Reports*, 19, 3-11.

Bartko, J. J. (1976). “On Various Intraclass Correlation Reliability Coefficients.” *Psychological Bulletin,* 83(5), 762-765.

Bartko, J. J. (1978). “Reply to Algina.” *Psychological Bulletin*, 85(1), 139-140.

Bernardin, H., and R. Beatty (1984). *Performance Appraisal: Assessing Human Behavior at Work*. Boston, MA: Kent Publishing Company.

Brown, R., and N. M. A. Hauenstein (2005). “Interrater Agreement Reconsidered: An Alternative to the rwg Indices.” *Organizational Research Methods*, 8(2), 165-184.

Cohen, S. L. (1978). “Standardization of Assessment Center Technology: Some Critical Concerns.” *Journal of Assessment Center Technology*, 1 1-10.

Cole, M., H. Field, and W. Giles. (2003). “What Can We Uncover About Applicants Based on Their Resumes?” *Applied HRM Research,* 8(2), 51-62.

Cole, M., R. Rubin, H. Field, and W. Giles (2007). “Recruiters’ Perceptions and Use of Applicant Resume Information: Screening the Recent Graduate.” *Applied Psychology: An International Review,* 56(2), 319-343.

Conway, J., R. Jako, and D Goodman (1995). “A Meta-Analysis of Interrater and Internal Consistency Reliability of Selection Interviews.” *Journal of Applied Psychology*, 80(5), 565-579.

Dipboye, R. L., B. Gaugler, T. Hayes, and D. Parke (2001). “The Validity of Unstructured Panel Interviews: More Than Meets the Eye?” *Journal of Business and Psychology*, 16, 35-49.

Donner, A., and G. Zou (2002). “Testing the Equality of Dependent Intraclass Correlation Coefficients.” *The Statistician*, 51(3), 367-379.

Gigone, D., and R. Hastie (1997). “Proper Analysis of the Accuracy of Group Judgments.” *Psychological Judgments*, 121(1), 149-167.

Griffin, M. (1989). “Personnel Research on Testing, Selection and Performance Appraisal.” *Public Personnel Management,* 18(2), 127-137.

Hagman, J. D. (1998). “Using Engagement Skills Trainer to Predict Rifle Marksmanship Performance.” *Military Psychology,* 10(4), 215-224.

Haynes, S., D. Richard, and E. Kubany (1995). “Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods.” *Psychological Assessment*, 7(3), 238-247.

Kumar, N., L. Stern, and J. Anderson (1993). “Conducting Interorganizational Research Using Key Informants.” *Academy of Management Journal*, 36(6), 1633-1651.

Lawshe, C. H. (1975). “The Quantitative Approach to Content Validity.” *Personnel Psychology*, 28, 563-575.

Latham, G. P., and K. N. Wexley (1982). *Increasing Productivity Through Performance Appraisal*. Reading, MA: Addison-Wesley.

Lu, L, N. Shara (2007). “Reliability Analysis: Calculate and Compare Intra-Class Correlation Coefficients (ICC) in SAS.” *Statistics and Data Analysis*, NESUG, <http://www.nesug.org/proceedings/nesug07/sa/sa13.pdf>.

McDaniel, M., F. Schmidt, and J. Hunter (1988). “A Meta-Analysis of the Validity of Methods for Rating Training and Experience in Personnel Selection.” *Personnel Psychology,* 41, 283-314.

McDaniel, M., D. Whetzel, F. Schmidt and S. D. Maurer. (1994). “The Validity of Employment Interviews: A Comprehensive Review and Meta-Analysis.” *Journal of Applied Psychology,* 79(4), 599-616.

McEvoy, G. M. (1990). “Public Sector Managers’ Reactions to Appraisal by Subordinates.” *Public Personnel Management*, 19, 201-212.

McKillip, J., and C. Cox (1998). “Strengthening the Criterion-Related Validity of Professional Certifications.” *Evaluation and Program Planning,* 21(2), 191-197.

Meng, Xiao-li, R. Rosenthal, and D.B. Rubin (1992). “Comparing correlated correlation coefficients.” *Psychological Bulletin*, 111(1), 172-175.

New York State Department of Civil Service (1976). *Guidelines for Training and Experience Evaluations.* Albany, N.Y.: OGS Printing Office.

Owens, W. (1976). “Background data.” In M. D. Dunnette (Ed.) *Handbook of Industrial and Organizational Psychology* (pp. 609-644). Chicago: Rand McNally.

Payton, M., M. Greenstone, and N. Schenker (2003). “Overlapping Confidence Intervals or Standard Error Intervals: What Do They Mean in Terms of Statistical Significance?” *The Journal of Insect Science*, 3 (34), 1-13, <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC524673/>.

Posthuma, R., F. Morgeson, and M. Campion (2002). “Beyond Employment Interview Validity: a Comprehensive Narrative Review of Recent Research and Trends Over Time.”  *Personnel Psychology,* 55(1), 1-81.

Pulakos, E., and N. Schmitt (1995). “Experience-Based and Situational Interview Questions: Studies of Validity.” *Personnel Psychology,* 48(2), 289-308.

Sackett, P., and M. Wilson (1982). “Factors Affecting the Consensus Judgment Process in Managerial Assessment Centers.” *Journal of Applied Psychology,* 67(1), 10-17.

Shrout, P., and J. Fleiss (1979). “Intraclass Correlations: Uses in Assessing Rater Reliability.” *Psychological Bulletin*, 86(2), 420-428.

Sneed, J., V. Vivian, and A. D’Costa (1987). “Work Experience as a Predictor of Performance: A Validation Study.” *Evaluation & the Health Professions,* 10(1), 42-57.

Tenopyr, M., and P. Oeltjen (1982). “Personnel Selection and Classification.” *Annual Review of Psychology*, 33, 581-618.

Voskuijl, O., and T. van Sliedregt (2002). “Determinants of Interrater Reliability of Job Analysis: A Meta-Analysis.” *European Journal of Psychological Assessment*, 18(1), 52-62.

Wagner, S., C. Rau, and E. Lindermann (2010). “Multiple Informant Methodology: A Critical Review and Recommendations.” *Sociological Methods & Research*, 38(4), 582-618.

Wanguri, D M. (1995). “A Review, an Integration, and a Critique of Cross-Disciplinary Research on Performance Appraisals, Evaluations, and Feedback: 1980-1990. *The Journal of Business Communication*, 32(3), July, 267-293.

Wernimont, P., and J. Campbell (1968). “Signs, Samples, and Criteria.” *Journal of Applied Psychology,* 52, 372-376.

Winer, B. J. (1971). *Statistical Principles in Experimental Design* (2nd ed.). New York: McGraw-Hill.