EVALUATING TRAINING AND EXPERIENCE: DO MULTIPLE RATERS OR CONSENSUS MAKE A DIFFERENCE?

Baugher, Dan
Pace University

Weisbord, Ellen
Pace University

Eisner, Alan
Pace University

ABSTRACT
The study considers 64 and 35 candidates seeking in-line promotions in a State Agency to financial analyst (FA) and upper management (UM) positions, respectively. Subject matter experts (SMEs) rated prior relevant experience using a training and experience exam (T&E). Three T&E rating approaches are contrasted: one rater, two raters, and two raters with hybrid consensus. Interrater reliability before hybrid consensus showed moderate agreement with an intraclass r of .75 and .74 for the FA and UM position, respectively. About half of the scores for each position were considered sufficiently divergent to merit a consensus meeting. Interrater reliability improved significantly ($p \leq .05$) with an intraclass r of .93 and .90 following hybrid consensus for the FA and UM position, respectively. A paired t-test showed no significant differences between the means from the three rating approaches ($p > .05$) suggesting no difference in leniency or harshness. The Pittman-Morgan t-test for comparing correlated variances showed that the variance from the three approaches did not differ significantly for the UM position ($p > .05$). For the FA position, the variability resulting from one rater was significantly greater than that for two raters ($t = 2.22$, $df = 62$; $p \leq .05$) and for two raters with hybrid consensus ($t = 2.54$, $df = 62$; $p \leq .01$) though the variability resulting from two raters did not differ significantly from that for two raters with hybrid consensus ($t = 1.32$, $df = 62$; $p > .05$).

INTRODUCTION

Prior training and experience is frequently used in the selection of job applicants in the public sector. This information is formally assessed, often by subject matter experts (SMEs), through the use of training and experience exams (T&Es). T&Es are structured application blanks where candidates for a position describe how their prior training and experience fits the knowledge, skill, and ability (KSA) requirements of the position for which they are applying. While data on the prevalence of the T&E is difficult to obtain, a 1979 survey of 11 localities and 16 states showed that T&Es were more frequently used than any other selection technique except for written tests (Beardsley, 1976; New York State Department of Civil Service, 1976).

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 instruments 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 an assessment of prior experience. The T&E is a variation on these techniques with the common goal of evaluating the quality and relevance of prior experience.

Research 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. Likewise, Sneed et al. (1987) found a positive relationship between the quality of prior work experience and the job performance of dietary managers.

The quality and relevance of prior experience has been found to relate to future job performance (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. 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 differences among different rating approaches for evaluating prior experience. The evaluation of prior experience is time-consuming and expensive. This is especially true for measures employing an essay format, as is often the case for managerial and other complex job positions. In addition, the reliability of alternative approaches sets an upper limit to the empirical validity that can be found 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

In examining the relationship of expert ratings to various criteria, a number of studies have found such ratings to have a positive correlation to the criterion of interest. 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, for the relationship between unstructured panel interviews by experts for training success and job performance of 513 correction officers. Expert ratings by the panels provided incremental value over two paper credentials for predicting officer training success and job performance.

KSA-based T&Es are typically rated by subject matter experts (SMEs). The number of raters varies, as do the methods for aggregating multiple ratings. Gigone and Hastie (1997) found that the unweighted means of individual judgments of an organization's properties outperformed behavioral approaches asking informants to work out their differences and reach a common response. This behavioral approach is often referred to as the "consensual" approach (Kumar, et al., 1993). For evaluations of an organization's properties where a single response for the organization is desired, Kumar, et al. (1993) developed a hybrid approach combining the averaging approach with the consensual approach for coming up with one common score. That approach specified that raters needed to go to consensus when a difference in ratings of 2 or more points on a 7-point scale occurred. In their study, consensus was needed for 15 percent of the questions, on average. Wagner, et al. (2010) suggest that the hybrid approach, which makes use of

consensus only for responses that differ by some specified value, results in less effort and is a very useful method for combining responses into a single organizational response.

When it comes to comparing outcomes, there is little empirical evidence on the relative performance of averaged ratings versus those obtained through full consensus or a hybrid approach to consensus (Wagner, et al., 2010). Multiple expert raters take time to make their evaluations and their use can be cost prohibitive. Also, when multiple raters are used, it is not uncommon for the individual ratings forming the basis of the composite rating of prior experience or within an oral exam, to be aggregated and no record of the individual ratings kept.

T&Es are typically developed and defended on the basis of content validity, which makes the need for such data less important than for other validation techniques. The defense of content valid instruments is based primarily on the rigor used in developing the elements assessed by the technique to assure they are relevant to and representative of a targeted construct (Haynes, et al., 1995), and less on the empirical results of the ratings. In contrast, empirical validation techniques that focus on the correlation of a given technique with other constructs or criteria focus on tracking the ratings made in each use of an instrument.

Multiple raters for performance appraisal have a long history. Compared to traditional single rater systems, research demonstrates greater benefits from multiple rater performance appraisals. 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. Multiple raters have also been shown to provide an improved legal defensibility over ratings by one person (Bernardin and Beatty, 1984). Multiple ratings of performance appraisals are frequently reviewed by the employee, in contrast to what occurs for T&E exams and for other rating processes used for employment selection including group interviews.

Multiple ratings also improve the accuracy of prior experience evaluations. By helping to assure that observed scores are subject to less random error, multiple ratings help 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 the behaviors they included, a single rater could reach an estimated reliability of .84 or greater. For KSA-based job analyses, interrater reliabilities of .80 were more difficult to achieve, needing from five to nine raters to reach that standard.

In their meta-analysis of selection interviews, Conway, et al. (1995) found similar results for 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. They also found that behavioral 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.

For assessment centers, Cohen (1978) believed consensus was a central part of the rating process. In contrast, Sackett and Wilson (1982) suggested that consensus may not be necessary. For 18 ratings made on 719 individuals they found that the use of a mechanical decision rule, in the absence of consensus, can predict consensus results and overall assessment center results with 75.0% and 94.5% accuracy, respectively. As was the case for multiple raters, consensus could improve perceptions by employees that a given rating process was carefully administered (McEvoy, 1990). Consensus ratings, like many group decisions, are more likely to be less extreme than individual ratings making an increase in harshness, leniency, or variability unlikely in comparison to what might occur from the use of one rater.

While job assessment is different from the assessment of prior experience, the research regarding KSA-based job analyses suggests that reaching the threshold value of .80 advocated by some for strong agreement (Brown & Hauenstein, 2005; Wagner, et al., 2010) may take two raters, not one. The need for multiple raters to reach this threshold value gets further support from the higher reliabilities found for panel interviews and multiple ratings in interview settings.

The support for consensus is less clear. For interviews, ratings combined behaviorally did not yield greater reliabilities than those combined mechanically. 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. In the development of single measures of an organization's characteristics based on multiple ratings, hybrid consensus has been advocated as better than full consensus for combining multiple informant ratings but the benefit of it for reliability coefficients is not clear.

PURPOSE OF THE STUDY

This study focuses on contrasting the results from three different approaches to rating training and experience as measured by a T&E. The approaches are (1) one expert rater, (2) two expert raters, and (3) two expert raters with hybrid consensus where T&E scores differing by more than a specified amount must go to consensus. For the two rater condition, T&E scores were developed by averaging the scores from the two raters. For the hybrid consensus condition, T&E scores were also developed by averaging the scores from the two raters but the scores averaged were post-consensus scores for those scores requiring consensus meetings and "no consensus" scores for those not requiring consensus. The goal of the study is to contribute to understanding of how multiple raters and hybrid consensus affect ratings of training and experience. The means and standard deviations for all three approaches are considered to determine if changes in leniency, harshness, or variability can be expected from the use of two raters or hybrid consensus. In addition, the impact of hybrid consensus on interrater reliability is evaluated.

HYPOTHESES

Three hypotheses are considered. H1 states that there will be no difference in the means for the T&E scores resulting from the three approaches. None of the research presented earlier suggests that multiple raters or hybrid consensus result in more harsh or more lenient ratings.

Hypothesis 2 states that the variability of T&E scores will be greater in the single rater condition. It goes on to state that greater variability will be present in the two rater condition compared to that resulting from the hybrid consensus condition. At the heart of H2 are concerns over multiple raters and consensus producing an averaging effect. For two ratings averaged together, the averaging can reduce variability if many divergent ratings exist. For hybrid consensus meetings, it is possible that raters will choose more often to move to the middle than to another's extreme rating thereby reducing variability. Both can be important concerns in the scoring of tests used for promotional decisions as more variability is generally preferable to less variability.

H3 proposes that the intraclass reliability for two raters will improve following hybrid consensus. This is advocated for the entire sample even though consensus was not required for all candidates since the primary concern lies in the reliability of the process. It is likely that reliability will improve as raters review divergent ratings and rethink their application of standards to such ratings. When this occurs, it reduces the random error that can easily occur as the result of confusion on how to apply standards for some KSAs to some kinds of prior experience.

The hypotheses are:

H1: The mean for T&E expert scores resulting from the one rater, two rater, and hybrid consensus conditions will not differ.

H2: The variability of T&E expert scores resulting from the one rater condition will be of a greater magnitude than those resulting from the two rater and hybrid consensus conditions. Also, the variability of T&E expert scores resulting from the two rater condition will be of a greater magnitude than those resulting from the hybrid consensus condition. These differences in variability will exist across all T&Es and for the subset of T&Es requiring consensus.

H3: The intraclass correlation coefficient for T&E expert scores resulting from hybrid consensus will be of a greater magnitude than the coefficient resulting from averaging the ratings of two raters, without consensus, to create all scores.

While this study considers two promotional positions, a Financial Analyst (FA) position and an upper management (UM) position, it is expected that the hypotheses will apply in the same manner to both positions.

SUBJECTS

Subjects were candidates for promotion to one of two positions within a State Agency. Employees who sought the financial analyst (FA) position wished to move from a "journeyman" to an "expert-level" functional position considered critical to the Agency's mission. Those applying for the upper management (UM) position were upper-middle level managers seeking an upper management position, the second highest management position within the Agency.

Both the incumbent and promotional position required the same KSAs and on-the-job behaviors but differed in their complexity and level of responsibility. In order to apply for promotion into either position, candidates had to be in the incumbent position for at least one year. This resulted in 64 subjects applying for promotion to the FM position and 35 subjects applying for promotion to the upper management position.

INSTRUMENTS AND RATING PROCESS

Prior training and experience was assessed through a KSA-based T&E exam. Following a traditional KSA-based development process, the exam measured the candidate's possession of KSAs important to performance in the promotional position. The T&E for the FA position contained 26 KSA items while the T&E for the UM position contained 19 KSA items. Both were developed through a rigorous methodology to assure that the KSAs were important in the promotional position (Lawshe, C., 1975; New York State Department of Civil Service, 1976). Since the promotions were in-line in nature, the same KSAs were also important to success in the incumbent position.

Candidates took the exam home to complete. Candidates could cite any past experience that matched the KSAs as long as it could be verified. They were not required to limit their response to experiences within the Agency. Responses were open ended in nature and were often from one to three pages in length. No page limit was set, as earlier research on the process showed that page length did not correlate with evaluations of prior experience. All candidates knew that their responses were subject to an audit to assure that they had engaged in the experiences as described on the T&E. For each candidate, three items were randomly selected for verification. If the contact listed could not confirm that an experience took place or was possible, all KSA items were verified. None in this study had verification difficulties.

SME raters for the FA and UM positions were selected from among middle managers (next grade or higher) and top managers within the Agency, respectively. In all cases, SMEs had to be familiar with both the incumbent and promotional positions. SMEs were not direct supervisors of or familiar with any candidate seeking promotion.

SMES were provided with a comprehensive training program. A scoring book was designed by SMEs within the Division to provide examples of responses for each of the four possible ratings for all KSA items. The four rating possibilities are shown below. Each SME was blind to the ratings of the other. When two raters' scores differed by more than seven points, a consensus meeting was required to bring the two scores within seven points of each other. This "rule of seven" resulted in 51.6% of the T&Es for the FA position and 45.7% for the UM position requiring consensus. Consensus meetings typically lasted about 45 minutes.

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 experience.

Since two SMEs evaluated each T&E, 128 FA T&Es and 70 UM T&Es had to be evaluated by the SMEs. For the FA position, eight SMEs rated a total of 16 T&Es each, in differing pair combinations. They were assigned from 5-11 T&Es to evaluate as a first or second rater depending on their availability and the candidates rated. For the UM position, six SMEs rated a total of 11 to 12 T&Es, in differing pair combinations. These SMEs were assigned to 4-8 T&Es as a first or second rater.

To minimize any potential order effect for the raters, the T&E score for the one-rater condition was randomly selected from the two ratings for each candidate. The Bernoulli distribution was employed to select the score from ratings made by rater 1 or rater 2. For the FA position, 47% of rater 1 scores and 53% of rater 2 scores were selected. For the UM position, 54% of rater 1 scores and 46% of rater 2 scores were selected.

For ease of comparison, T&E scores from all three conditions were normalized to range from 0 to 100. For the one rater condition, this score was derived from the ratings of one randomly selected rater. For the two-rater condition, this score was the average for the two raters. For the hybrid consensus condition, this score was also the average for the two raters. For those scores requiring hybrid consensus, it was the average of the post-consensus scores. For those scores not going to consensus, it was the average of the "no consensus" scores for the two raters.

INTER-RATER AGREEMENT AND RATER RELIABILITY

Inter-rater agreement on individual items was assessed by using Finn's r (r_f), which can be interpreted as the proportion of correspondence in the observed ratings that is not due to chance (Finn, 1970). For the FA T&E, Finn's r ranged from .725 to .845 with an item average of .783. For the UM T&E, Finn's r ranged from .65 to .825 with an item average of .731. After hybrid consensus, Finn's r ranged from .819 to .936 with an item average of .886 for the FA T&E and from .725 to .923 with an item average of .818 for the UM T&E. The coefficients before and after hybrid consensus are quite respectable, especially in light of the essay format of the information provided.

In the assessment of H3, 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 a one-way ANOVA intraclass correlation or ICC(1). Choice of an intraclass correlation is largely a function of how raters are used in a study. In this study, the same pair of raters did not rate each candidate. Since raters cannot be a factor, the

one-way ANOVA is appropriate (Bartko, 1976; Bartko, 1978; Shrout and Fleiss, 1979). ICC(1) can be interpreted as indicating the level of intraobserver consistency one can expect should the same background be evaluated using the same coding scheme with observers of equivalent training in the future.

The intraclass correlation for the reliability of average ratings is also used in testing H3. 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 correlations considered in H3 are correlated correlation coefficients because the same subjects are rated in all three conditions. This made evaluation of the statistical significance of their difference far from straightforward. 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, nor easy to perform, and their power needs further investigation.

As a result, H3 is evaluated in this study by comparing the overlap in the 95% confidence intervals for the coefficients before and after hybrid consensus. In this approach, the absence of any overlap is a strong indicator that the coefficients differ (Lu and Shara, 2007; Payton, et al., 2003). Some would 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, using an 83% to 84% size for the intervals gives an alpha of .05. In evaluating H3, the 95% interval is used to test the hypothesis that the intraclass correlation for the study increased following hybrid consensus. The 84% interval is used when the more conservative 95% interval failed to show a statistically significant difference and the standard errors were sufficiently similar to permit the use of this approach.

RESULTS

Hypothesis 1 is supported for both job titles. H1 predicted that T&E Expert mean scores would not differ as a function of rating approach. Table 1 provides the means for the three rating approaches for all T&E scores and for scores going to hybrid consensus. The paired t-test was used to test the differences and none are statistically significant.

For the FA position, the t values for one rater by two raters, one rater by hybrid consensus, and two raters by hybrid consensus are $t = .45$, $t = .86$, $t = 1.27$ ($df = 63$, $p > .05$), respectively. For the UM position, the t values for one rater by two raters, one rater by hybrid consensus, and two raters by hybrid consensus are $t = .22$, $t = .32$, and $t = .59$ ($df = 34$, $p > .05$), respectively. For those going to hybrid consensus for the FA position, the t values for one rater by two raters, one rater by hybrid consensus, and two raters by hybrid consensus are $t = .73$, $t = 1.14$, and $t = 1.28$ ($df = 32$, $p > .05$), respectively. For T&E scores going to hybrid consensus for the UM position, the t values for one rater by two raters, one rater by hybrid consensus, and two raters by hybrid consensus are $t = .87$, $t = .91$, and $t = .58$ ($df = 15$, $p > .05$), respectively.

While the trends suggest that higher averages and leniency may be a possibility in the move from one rater to two raters and from two raters to hybrid consensus, the differences were not statistically significant. Typical of smaller sample studies, the power to discover the small differences uncovered here ranges from about 20-25%. Power analysis suggests that it would take from 150-200 subjects to have a power of 70% for detecting a 2 point difference at the .05 alpha level for similar sets of data if, in fact, that difference was present. It would take many more subjects to reach a power of 70% to discover these

smaller differences. Given the relatively small changes seen, it seems reasonable to accept that the null position of H1 is supported.

Table 1
Mean, Standard Error, and Standard Deviation for the Three T&E Rating Approaches

Position	Sample		One Rater	Two Raters	Hybrid Consensus
FM Position	All	Mean	52.6	53	53.4
	n=64	S.E.	2.28	2.07	2.02
		S.D.	18.3	16.6	16.2
	Consensus	Mean	50.9	52	52.7
	n=33	S.E.	3.09	2.59	2.44
		S.D.	17.8	14.9	14.0
UM Position	All	Mean	59.2	59.5	59.7
	n=35	S.E.	3.80	3.53	3.55
		S.D.	22.5	20.9	21.0
	Consensus	Mean	57.5	60	60.4
	n=15	S.E.	6.78	5.99	6.06
		S.D.	27.1	24	24.2

Hypothesis 2 is partially supported for the FA position but not for the UM position. H2 predicted greater variability for the single rater condition compared to the other conditions and for the two-rater condition compared to the hybrid consensus condition. Since these are variances from correlated variables, the Pitman-Morgan test of significance was applied to the variances. This yields a t statistic with N-2 degrees of freedom (see Kenny, 1953). The more conservative two-tailed test was used to test the variance differences since H2 was not based on a strong base of research. Table 1 also provides the standard deviations for the three rating conditions for all T&E scores and for scores going to hybrid consensus.

For all FA T&E scores, greater variability was shown in the one-rater condition as compared to the two-rater condition ($t = 2.21$, $df = 62$, $p \leq .05$) and the hybrid consensus condition ($t = 2.55$, $df = 62$, $p \leq .01$). No difference in variability was found between the two-rater and the hybrid consensus condition ($t = 1.33$, $df = 62$, $p > .05$). Similar results were found in the FA position for the subset of scores going to hybrid consensus. Greater variability was shown in the one-rater condition as compared to the two-rater condition ($t = 2.11$, $df = 31$, $p \leq .05$) and the hybrid consensus condition ($t = 2.59$, $df = 31$, $p \leq .01$). No

difference in variability was found between the two-rater and the hybrid consensus conditions ($t = 1.51$, $df = 31$, $p > .05$).

For all UM T&E scores, no significant difference in variability was found for the different rating conditions. The t values for one rater by two raters, one rater by hybrid consensus and two raters by hybrid consensus are $t = 1.17$, $t = .95$, and $t = .36$, ($df = 33$, $p > .05$), respectively. Likewise, no significant difference in variability was found for the subset of scores going to hybrid consensus. The t values for one rater by two raters, one rater by hybrid consensus, and two raters by hybrid consensus are $t = 1.07$, $t = .87$, and $t = .28$ ($df = 13$, $p > .05$).

The partial support for H2 lies in the greater magnitude of variability for the one-rater condition as compared to the two-rater and hybrid consensus conditions for the FA position. As hypothesized, this occurred across all T&E scores. It also occurred for the subset of T&E scores requiring a hybrid consensus meeting. While similar trends appeared for the UM position, the differences were not statistically significant. This may be partially due to the smaller sample size studied for that position. The higher variability anticipated by H2 for the two-rater condition over that for the hybrid consensus condition did not materialize for either job position or for the subset of T&E scores requiring hybrid consensus.

While not part of any hypothesis including H1 and H2, the means and variability for scores going to hybrid consensus were contrasted with those not requiring consensus. The independent t-test was used to check the differences for these two sets of scores in the one-rater condition. The differences were not statistically significant for the FA position ($t = .754$, $df = 62$, $p > .05$) or the UM position ($t = .397$, $df = 33$, $p > .05$). The trend toward greater variability for T&E scores requiring consensus is a bit stronger, as shown in Table 1. The Levene Test for equality of independent variances did not show significant differences for the FA position for the one rater condition ($F = .212$, $df = 62$, $p > .05$), two-rater condition ($F = 1.38$, $df = 62$, $p > .05$), or the hybrid consensus condition ($F = 2.05$, $df = 62$, $p > .05$). No significant differences in variability existed for the UM position under the same conditions. With a df of 33, the Levene test showed $F = 3.49$ ($p > .05$), $F = 3.29$ ($p > .05$), and $F = 2.81$ ($p > .05$), respectively. With greater power from a larger sample size, these differences in variability may have shown statistical significance for the UM position but likely not for the FA position.

Hypothesis 3 is supported. H3 predicted that interrater reliability for the T&E scores would benefit significantly from the hybrid consensus meetings. The ICC(1) for the FA position was .75 prior to hybrid consensus and .93 following hybrid consensus. Figure 1 shows the 95% confidence intervals for the intraclass correlation before (B) and after hybrid consensus (A). There is no overlap between the two intervals indicating that they are significantly different ($p \leq .05$). As neither interval includes a 0, it is also clear that they are statistically significant correlations ($p \leq .05$). Since this conservative test of the statistical significance of the differences showed no overlap, the 84% intervals are not presented.

The ICC(1) for the UM position was .74 before hybrid consensus (B) and .90 after hybrid consensus (A). Figure 2 shows the 95% confidence intervals for the intraclass correlation before (B) and after hybrid consensus (A). There is a small degree of overlap between the intervals as shown. With this conservative test, the difference does not reach statistical significance. For this reason, the 84% intervals are shown in Figure 3. As Table 1 shows, the standard errors for the means are close and the means are not significantly different allowing for this more liberal test. With this test, the intervals do not overlap, suggesting that the differences are statistically significant ($p \leq .05$). It is likely that the more conservative test, which used 95% intervals, was unable to reach significance given the small sample size involved. Neither interval includes a 0 indicating that both correlations are statistically significant ($p \leq .05$).

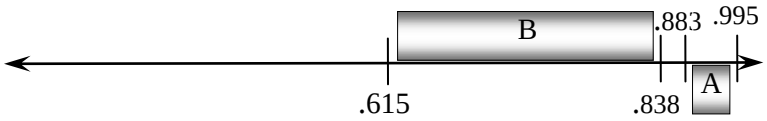

Figure 1. 95% Critical Region for Intraclass Correlations: FA Position.

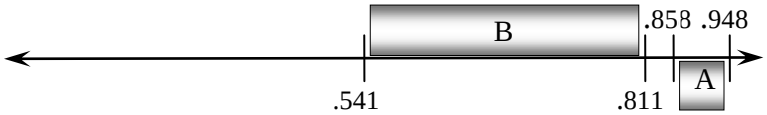

Figure 2. 95% Critical Region for Intraclass Correlations: UM Position.

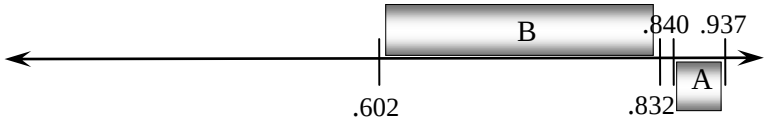

Figure 3. 84% Critical Region for Intraclass Correlations: UM Position.

ICC(2) was also computed before and after hybrid consensus for the FA and UM position. This correlation 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 hybrid consensus was .86 and .96 following hybrid consensus. As was true for the ICC(1) for this position, the 95% confidence intervals for the correlations do not overlap. They are .762 to .912 before hybrid consensus and .938 to .977 following hybrid consensus, suggesting a significant difference ($p \leq .05$).

For the UM position, ICC(2) before hybrid consensus was .85 and .95 following hybrid consensus. The 95% confidence intervals for the correlations overlap somewhat as did the ICC(1) intervals for this position. They are .702 to .923 before hybrid consensus and .895 to .973 following hybrid consensus. As Table 1 shows, the standard errors of the means are close and the means are not significantly different. Again, use of the less conservative 84% confidence intervals to compare the difference in the correlations resulted in no overlap showing a statistically significant difference ($p \leq .05$). These intervals are .752 to .908 before hybrid consensus and .913 to .968 after hybrid consensus suggesting a significant difference ($p \leq .05$).

While not a part of H3, the improvements shown for ICC(1) for the subset of scores going to hybrid consensus were also compared. About half of the T&Es in both the FM and UM position were sent to a consensus meeting. For the FA position, ICC(1) was .53 before hybrid consensus and .90 after hybrid consensus for T&E scores requiring consensus with 95% confidence intervals of .230 to .733 before hybrid consensus and .830 to .936 after hybrid consensus. The absence of overlaps between these confidence intervals indicates that the correlations are significantly different ($p \leq .05$). The absence of a 0 in the intervals indicates that the correlations are statistically significant ($p \leq .05$).

For the UM position, ICC(1) was .64 before hybrid consensus and .89 following hybrid consensus for T&E scores requiring consensus with 95% confidence intervals of .234 to .854 before hybrid consensus and .715 to .959 after hybrid consensus. There is overlap between the intervals. Use of the 84% confidence intervals also showed some overlap with intervals of .361 to .811 before hybrid consensus and .777 to .946 after hybrid consensus. This overlap is likely due, in part, to the small sample of 15 scores included in the comparison. Also, there is less change in interrater reliability for scores requiring hybrid consensus for this position than for the FA position. The absence of a 0 in the intervals indicates that the correlations are statistically significant ($p \leq .05$).

DISCUSSION

This study investigated the complex rating process of evaluating prior experience against KSA-based standards when information is collected in essay format. A T&E examination was used for evaluation, which is a common practice in State government. The possibility of the means and variability differing under conditions of one rater, two raters and hybrid consensus was examined. Also, interrater reliability resulting from scores developed from the average of two ratings and from a hybrid consensus approach were compared.

The intraclass correlations show that SME ratings of prior experience for both a financial analyst (FA) position and an upper management (UM) position resulted in moderate agreement among the raters without hybrid consensus meetings. The correlations of about .75 for both positions suggest that raters can make moderately consistent ratings of KSA-related prior experience from essay answers with a remarkable degree of consistency. When hybrid consensus was implemented, interrater reliability for the process increased considerably, moving from moderate levels of agreement to high levels of agreement yielding intraclass correlations of .90 or greater for the two positions.

As is always the case, the intraclass correlation for the average of two raters was higher before and after hybrid consensus than it was for single ratings. This is because average scores tend to be more stable than single scores. Nevertheless, this coefficient also improved before and after hybrid consensus, moving from about .85 to .95 for the two positions. This finding is relevant to those who expect to use the average of two ratings as opposed to a single rating in such a setting.

Some T&E scores benefited from the additional scrutiny of hybrid consensus meetings. Raters came together in their assessments as a result of these meetings. The need for consensus for some scores is not surprising. Individual raters have personal views about the value of prior experience, and no scoring rubric can take into consideration the full array of experience that might be documented in such a process.

The use of two raters and hybrid consensus did not result in increased leniency or harshness for the process as a whole or for the T&E scores going to consensus. Changes in variability did occur, but not completely as expected. The scores generated by one rater tended to have greater variability than scores generated by the average of two raters or hybrid consensus. The tendency for variability to decrease when two raters were involved, whether by two-rater average or hybrid consensus, was most evident and statistically significant for the FA position. It did not reach statistical significance for the UM position.

Unexpectedly, hybrid consensus did not reduce variability further than two-rater averaging. While not hypothesized, the absence of a reduction in variability is an important outcome. When variability shrinks, small differences in scores can lead to big differences in outcomes, much the way that reduced variability in a final exam can cause small differences in exam scores to lead to big differences in the grade assigned. While the absence of a lower variance for the hybrid consensus meeting in comparison to the two rater condition did not support H2, this is a positive outcome for the rating and scoring process as shrinkage of variability is rarely beneficial for a host of psychometric reasons.

The most heartening result shown by this study lies in the remarkable ability of busy managers to evaluate complex write-ups of prior experience with, at minimum, a moderate degree of agreement. The 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 a moderate degree of reliability can be achieved, though prior experience that does not easily fit the scoring rubric or rater views of quality will likely be less reliably scored. Hybrid consensus was useful in this setting in that it significantly increased interrater reliability, thereby increasing the maximum empirical validity that could be found for the T&E measure of prior experience.

That said, this study looked only at improvements in reliability and outcomes due to multiple raters and hybrid consensus. Whether there is a concomitant improvement in the relationship of prior experience to job performance is a different, and important, research question. The authors plan to conduct future research on this question, using additional T&E scores and including performance ratings for those taking the exams.

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.

Barko, 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.

Beardsley, V. A. (1976). *A Study of the Rating of Education and Experience as an Examination Method in the Pennsylvania State Civil Service Commission.* Harrisburg, Pa.: Pennsylvania State Civil Service Commission.

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 r_{wg} Indicies." *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.

Finn, R. H. (1970). "A Note on Estimating the Reliabiity of Categorical Data." *Educational and Psychological Measurement*, 30(1), 71-76.

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 an 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.

Kenny, D. T. (1953). "Testing of Differences Between Variances Based on Correlated Variates." *Canadian Journal of Psychology*, 7(1), 25-28.

Kumar, N., L. Stern, and J. Anderson (1993). "Conducting Organizational 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.

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 Jounral 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.