Personnel Assessment and Decisions

Volume 1 | Issue 1 Article 3

2015

The Validity of Individual Psychological Assessments for Entry-Level Police and Firefighter Positions

Ilianna H. Kwaske
Illinois Institute of Technology

Scott B. Morris
Illinois Institute of Technology

Recommended Citation
Kwaske, Ilianna H. and Morris, Scott B. (2015) "The Validity of Individual Psychological Assessments for Entry-Level Police and Firefighter Positions," *Personnel Assessment and Decisions*: Vol. 1 : Iss. 1 , Article 3.
DOI: https://doi.org/10.25035/pad.2015.003
Available at: https://scholarworks.bgsu.edu/pad/vol1/iss1/3

The Validity of Individual Psychological Assessments for Entry-Level Police and Firefighter Positions

Iliana H. Kwaske[1] and Scott B. Morris[1]

1. Illinois Institute of Technology

KEYWORDS

individual psychological assessment, validity of individual psychological assessment, individual assessment for employee selection, hierarchical linear models for validation

ABSTRACT

Although individual psychological assessment is widely used in employee selection, the empirical research on the validity of individual assessments is sparse. A multistage, multisite study examined the validity of individual assessments for police officer and firefighter positions. Results showed that assessor judgments were largely unrelated to standardized test results and that both assessor judgments and standardized tests were only weakly related to job performance ratings. Differences in validity across assessors were also found, with some assessors providing better predictions than others.

Individual psychological assessment (IPA) is a widely used method of evaluating abilities, personality, and person–job fit of job applicants (Prien, Schippmann, & Prien, 2003). However, unlike other selection tools, the research on IPAs has been sparse. For more than 50 years, researchers have been calling for more empirical evidence to support the use of IPAs for personnel selection (e.g., Hilton, Bolin, Parker, Taylor, & Walker, 1955). The current study answers the call for research by conducting an in-depth validation of one consulting firm's IPA process.

Individual Psychological Assessment

IPA is a process of collecting information regarding a job candidate's knowledge, skills, abilities, and personality through the use of individually administered selection tools, including standardized tests and interviews, and integrating this information to make an inference regarding the individual's suitability for a particular position (Jeanneret & Silzer, 1998). What distinguishes IPA is the use of assessor expert judgment to provide insight into the person as a whole (Highhouse, 2002). IPAs are often designed to evaluate a *combination* of knowledge, skills, abilities, and other attributes of the potential employee. Further, through personal interaction, the assessor may be able to observe subtle behavioral cues and interpret specific responses in the context of broader behavioral patterns (Silzer & Jeanneret, 2011), thereby providing information not available through standardized tests.

Despite these potential advantages, the empirical evidence in support of IPAs is limited. A review by Prien et al. (2003) identified 20 criterion-related validation studies, with most showing at least modest validity. More recently, a meta-analysis of 39 validation studies (Morris, Daisley, Wheeler, & Boyer, 2015) yielded an average validity (corrected for criterion unreliability) of .30 (.21 for nonmanagerial jobs). Both reviews reported considerable variability in validities across studies and noted that much of the available research was conducted before 1970.

Because individual assessments involve an assessor interpreting a battery of tests, it is important to examine the contribution of both the assessor and the test battery itself. A typical assessment involves administration of several standardized personality and cognitive ability tests, along with an interview and biodata form (Ryan & Sackett, 1987, 1998). Extensive research has found that these methods are all useful predictors of job performance (Schmidt & Hunter, 1998). In the context of public safety occupations, Hunter & Hunter (1984) reported a corrected validity of .42 for general cognitive ability. Subsequent research has reported similar validities for firefighters (Barrett, Polomsky, & McDaniel, 1999; Henderson, 2010) but lower validity for police officers (.27; Aamodt, 2004). Personality traits, in particular conscientiousness, have been found to be

Corresponding author:
Iliana H. Kwaske
Email: ihk@kwaske.com

useful predictors of police performance (corrected r = .22; Aamodt, 2004; Barrick & Mount, 1991), although we were unable to locate any research on the validity of personality for firefighters.

Thus, the typical IPA will assess for a broad array of job-relevant information about the candidate, and it is not surprising that the results of the assessment are predictive of job performance. What is less clear is the extent to which this validity is driven wholly by the tests themselves or whether the interpretation of test scores by an expert assessor adds to (or detracts from) this validity.

In understanding the role of assessors in IPA, it is useful to look at the process by which assessors form dimension ratings and recommendations from the test battery. Research by Ryan and Sackett (1987) found considerable discrepancies in IPA practices across assessors, as well as low interrater reliability in assessor judgments. Idiosyncratic interpretations of test data could produce inconsistencies across assessors that might weaken validity when recommendations come from different assessors (O'Brien & Rothstein, 2011). Thus, an evaluation of IPA should consider the relationship between test components and assessor ratings.

Whether assessors add to or detract from IPA validity is a matter of considerable debate (cf. Highhouse, 2008; Silzer & Jeanneret, 2011). Decades of research have demonstrated that a simple weighted average of test scores is generally as accurate, and sometimes more accurate, than clinical predictions (Ægisdóttir et al, 2006; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Kuncel, Klieger, Connelly, & Ones, 2013). Clinical judgment is a defining characteristic of IPAs, relying on the experience and skill of the assessor to interpret and integrate information gathered during the assessment (Weiner, 2003). Inaccuracies in clinical predictions could result from reliance on heuristic rules wherein assessors do not optimally use the available information (Camerer & Johnson, 1991; Garb, 2003). There has been limited empirical research on the incremental validity of assessor recommendations in IPA, with a few studies reporting validities as high as .77 (Meyer, 1956) or as low as -.05 (Miner, 1970), and studies failing to find incremental validity over the tests used in the assessment process (Holt, 1958; Huse, 1962; Meyer, 1956; Trankell, 1959). Empirical support for the use of individual psychological assessment is limited for a number of reasons, to include range restriction, criterion contamination, small sample size, unreliability of the predictor and criterion, and rater errors.

Given the role of assessor judgment in interpreting test results, another factor to consider is whether IPA validity is assessor specific. Assessors are not necessarily interchangeable; some are likely to prove better than others. Silzer and Jeanneret (2011) outline a number of essential competencies for assessors that might impact the usefulness of their ratings. Thus, it is not sufficient to validate the assessment tools; validity evidence is also needed for the judgment of each assessor (Morris, Kwaske, & Daisley, 2011).

Multilevel Analysis

IPA is most often used in situations with a small number of candidates and job openings. Therefore, in order to obtain a sufficient sample size for validation, it is useful to pool data across jobs or work settings. Similarly, IPA validation studies also typically combine data from multiple assessors.

Combining data across work sites or assessors creates nested data, where job applicants can be grouped by site or assessor. Consequently, observations at the individual level may not be independent, violating a key assumption of most statistical analyses. It is exactly this type of data that multilevel analyses were developed to model.

Hierarchical linear modeling (HLM; Raudenbush & Bryk, 2002) allows the creation of prediction models at multiple levels of aggregation. At Level 1, a multiple regression model represents the relationship between assessment data and job performance for individual candidates. The Level 1 model is estimated separately for each work site and assessor. The Level 2 model treats work site/assessor as the unit of analysis and estimates the mean and variance of the Level 1 regression coefficients across settings. This provides an overall estimate of the contribution of each predictor, as well as a means to test the extent to which these relationships differ across settings or assessors.

Research Questions

In summary, the existing literature paints an uncertain picture regarding the usefulness of IPA for the prediction of job performance. Although there is evidence that assessor ratings prediction job performance (Morris et al., 2015), questions remain concerning the contribution of assessors to IPA validity and whether the validity differs across assessors. To address these questions, the current study undertook a comprehensive validation on one consulting firm's assessment process for police officer and firefighter selection.

Research Question 1: What drives the assessor recommendations? Which components of the assessment battery are given the most weight in determining recommendations?

Research Question 2: Do assessors add value added beyond the assessment components in predicting job performance?

Research Question 3: Are there differences in IPA validity across assessors?

METHOD

Sample

Data on IPAs were obtained from archival records at a single consulting firm. Assessments were performed by six assessors with different levels of education, backgrounds (e.g., industrial organizational psychology, clinical psychology) and amount of experience.

Applicants for entry-level police officer or firefighter positions were sent for preemployment IPA between 1992 and 2001 by various municipalities in a large metropolitan area in the Midwest. Prior to the IPA, the candidates had already been prescreened through the municipalities' selection process, which could include mental ability/situational tests, board interviews, and physical agility testing. Complete assessment reports were available on 1,639 applicants from 39 municipalities (1,175 candidates from 26 police departments and 464 candidates from 13 fire departments). Analyses involving the prediction of job performance were based on 505 incumbents (360 police officers and 145 firefighters) for whom usable performance evaluations were obtained. Demographic information was not available.

Measures

All candidates completed the same battery of cognitive ability and personality tests, and a personal history form. The data available for this study consisted of scores on the personality and cognitive ability tests, as well as the assessment reports. Information specific to the interview was not recorded, except as it was reflected in the assessment report, and could not be separated from the other components of the assessment. Therefore, the interview was not included as a predictor in the current study.

The assessment protocol was developed through a job analysis conducted on police and firefighter positions. This analysis resulted in a competency model reflecting the characteristics critical for success in these positions. This competency model served as the basis for selecting the components of the assessment battery and as a framework for the assessor dimension ratings and the measure of job performance described below.

The competency models for police and firefighter jobs showed a high degree of overlap, although there were some differences. Given this overlap, the content of the assessment battery was the same for the two positions, and assessment ratings shared eight of nine dimensions in common. This study focused only on those common elements.

NEO PI-R. The *NEO PI-R* (Costa & McCrae, 1991) is a self-administered personality questionnaire containing 240 items covering the five domains of neuroticism, extraversion, agreeableness, openness to experience, and conscientiousness. Internal consistency coefficients for the domain scales range from .86 to .95 (Costa and McRae, 1991).

FIRO-B. The *Firo-B* (Consulting Psychologists Press, Inc., 1989; Hammer & Schnell, 2000) is a 54-item questionnaire designed to assess interpersonal needs. The scales measure an individual's needs for inclusion, control, and affection, each in terms of both the desire to express a behavior and the extent to which the characteristic is wanted from others. Coefficients of stability for the scales range from .71 to .82 (Hammer & Schnell, 2000). Only the control scale was used in the analyses because this construct was the most distinct from the constructs measured by the *NEO PI-R.*

Wesman Personnel Classification Test (WPCT). This test measures an applicant's verbal reasoning and quantitative ability (20 items each; Wesman, 1965). Reliabilities for the verbal reasoning and numerical ability scales range from α = .78 to .92.

Watson-Glaser Critical Thinking Appraisal (WGCTA). This test was designed to measure an individual's logical reasoning and critical thinking skills (Psychological Corporation, 1980). Split-half reliabilities range from .96 to .85, and test–retest reliability is .73 (Psychological Corporation, 1980).

Assessor dimension ratings. At the end of the IPA process, the psychologist integrated and synthesized the information gathered about the candidate (including the tests described above, a personal history form and a semistructured interview). The assessor's judgment was summarized through a narrative report and ratings on eight dimensions derived from a job analysis of public safety positions. Ratings were made on a four-point scale: poor, marginal, acceptable, and strong.

A principal component analysis revealed that seven of the eight dimensions (all but job preparation) loaded on a single factor. Therefore, these seven dimensions were averaged to form a single score labeled General Impressions (α = .87). Job preparation was maintained as a separate variable.

Overall recommendation. At the end of the individual assessment report, the psychologist provided the candidate's overall rating on a four-point scale ranging from 1 = *poor* to 4 = *strong*.

Job performance. Job performance data were collected for the purpose of this research in 2002, ensuring that all individuals had been on the job for at least one year. An employee performance form was designed by the consulting firm specifically for this study, and contained seven items derived from the same competency model that served as the framework for the assessment report. Structuring the performance measure to align with the assessment structure emphasizes the prediction of those aspects of performance that are most relevant to the purpose of the assessment (Pulakos Borman, & Hough, 1988) rather than prediction of overall job performance. All ratings were on a five-point scale.

A principal components analysis supported two performance dimensions. The first component, work disposition, comprised public service orientation, work attitude, conscientiousness, and adaptability (α = .89/.91 in firefighter/police samples). The second component, social/emotional competence, was comprised of interpersonal skills, openness to authority, and managing stress (α = .77/.81 in firefighter/police samples).

RESULTS

Descriptive statistics are reported in the Appendix. All variables were approximately normally distributed.

Correlates of Assessor Judgments

To understand the factors influencing assessor judgments, we correlated the components of the test battery with each of the three assessor ratings. The pooled within-organization correlations were computed separately for police and firefighter candidates and are summarized in Table 1.

Among police candidates, weak correlations were observed between the assessor general impression rating and several personality traits; the largest of these being with neuroticism (-.23) and agreeableness (.16). Notably, none of the ability measures were related to assessor recommendations.

The results were similar for firefighter candidates. Several personality traits correlated weakly with the general impression rating, and none of the ability measures correlated with assessor ratings. The strongest correlations were with conscientiousness (.26) and agreeableness (.22). Again, no meaningful correlations were found with the ability measures.

The job preparation rating showed little correlation with the standardized test scores. This is not surprising given that this measure reflected candidate experience, which was assessed through the interview and personal history form rather than the standardized tests.

Given the structure of the assessment process (i.e., assessors make dimension ratings and then an overall recommendation), we expected the overall recommendation to overlap substantially with the dimension ratings. This was indeed the case. The overall recommendation was strongly related to the general impression rating (r = .83 for police and r = .82 for firefighters) and to a lesser extent the job preparation rating (r = .33 for police and r = .30 for firefighters). Further, the overall recommendation showed a pattern of correlations with personality and ability tests that was quite similar although generally weaker than the correlations found for the general impression ratings.

Taken together, the set of tests accounted for only a small portion of the variance in the assessor judgments (9% for general impression, 3% for job preparation, and 6% for the overall recommendation). Thus, assessor ratings were only marginally influenced by the results of the standardized tests. Further, among the tests, assessors appeared to place more weight on personality measures relative to ability test results.

TABLE 1.
Correlations Between Assessment Tests and Assessor Judgments

	Police			Fire		
Predictor	GI	JP	OR	GI	JP	OR
Neuroticism	-.23	-.10	-.21	-.14	.04	-.11
Extroversion	.12	.04	.06	.11	.11	.18
Openness to experience	.01	-.08	-.02	-.01	-.03	-.04
Agreeableness	.16	-.05	.12	.22	.03	.20
Conscientiousness	.12	-.09	.02	.26	-.04	.20
Firo Control Expressed	.00	.04	-.03	-.06	.07	-.07
Firo Control Wanted	-.01	.01	-.05	.05	.12	.03
Watson Glaser	.05	.02	.09	.02	-.11	.01
Wesman Quantitative	-.02	-.02	.00	-.04	-.10	-.02
Wesman Verbal	.02	.00	.07	-.07	-.13	-.05
Assessor general impression rating	--	.23	.83	--	.26	.82
Assessor job preparation rating	--	--	.32	--	--	.30

Note. Pooled within-organization correlation across 26 police departments (N = 1,175) and 13 fire departments (N = 464). $|r| > .06$ for police and $|r| > .09$ for firefighters were significant at the .05 level. GI = assessor general impression rating. JP = assessor job preparation rating dimension. OR = assessor overall recommendation.

Prediction of Job Performance

The criterion-related validity of the IPA was analyzed using a multilevel hierarchical regression analysis, first entering the standardized test scores and then adding assessor judgments to the model in a second step. The incremental validity due to the assessor was assessed through the additional variance accounted for in the second step, as well as the standardized regression coefficients for each assessor rating.

The analyses were conducted using HLM6 (Raudenbush, Bryk, Cheong, & Congdon, 2004). A cross-classified random effects model was specified with assessor and municipality as random effects. The Level 1 model consisted of the individual test scores and assessor recommendations as predictors of job performance. The Level 2 model consisted of random effects for each municipality and assessor, as well as fixed effects representing the intercept and slope differences between police and firefighter positions. Separate analyses were conducted for each dimension of job performance.

In a multilevel model, the variance accounted for by a set of predictors (R^2) can be determined by the reduction residual variance relative to a baseline model. Because there are multiple sources of variance (examinees, municipalities and assessors), we defined R^2 in terms of the reduction in total variance across the three components (LaHuis, Hartman, Hakoyama, & Clark, 2014). For this analysis, a fixed slopes model was used where the slope of each predictor was constant across groups, and only the intercepts were allowed to vary across municipalities and assessors. The results are summarized in Table 2.

For the first performance dimension (work disposition), we first regressed performance onto the battery of standardized tests, accounting for 4% of the variance in performance. Among the tests, neuroticism was a significant predictor for both positions, although the strength of the relationship was not strong, with a standardized regression coefficient (β) of .10. Conscientiousness was found to interact with position, showing a positive relationship with performance for firefighters (β = .13) but a negative

TABLE 2.
Hierarchical Multilevel Regression Analysis (Unstandardized Regression Coefficients) for the Assessment Battery (Model 1) and Assessor Ratings (Model 2) as Predictors of Job Performance Dimensions

	Work disposition				Social/emotional competence			
	Model 1		Model 2		Model 1		Model 2	
Predictor	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE
Intercept	3.83	0.06 **	3.82	0.05 **	3.76	0.05 **	3.77	0.06 *
Intercept x Position	-0.08	0.12	-0.06	0.12	0.04	0.12	0.04	0.12
Neuroticism (N)	0.10	0.05 *	0.11	0.05 *	0.05	0.05	0.06	0.04
N x Position	-0.02	0.11	-0.01	0.11	0.13	0.10	0.14	0.10
Extraversion (E)	0.09	0.05	0.08	0.05	0.12	0.05 *	0.11	0.05 *
E x Position	0.07	0.11	0.07	0.11	0.13	0.11	0.12	0.10
Openness (O)	-0.05	0.04	-0.06	0.04	-0.06	0.04	-0.07	0.04
O x Position	0.04	0.09	0.04	0.09	-0.06	0.08	-0.04	0.08
Agreeableness (A)	0.00	0.05	0.00	0.05	-0.04	0.05	-0.05	0.05
A x Position	0.13	0.11	0.10	0.11	0.15	0.11	0.14	0.11
Conscientiousness (C)	-0.01	0.05	-0.03	0.05	0.00	0.04	-0.02	0.04
C x Position	-0.20	0.10 *	-0.22	0.10 *	-0.22	0.10 *	-0.20	0.10 *
Firo Control Expressed (CE)	-0.01	0.02	-0.01	0.02	-0.01	0.02	-0.01	0.02
CE x Position	0.02	0.04	0.02	0.04	0.00	0.04	0.00	0.04
Firo Control Wanted (CW)	-0.02	0.02	-0.02	0.02	-0.05	0.02 *	-0.05	0.02 *
CW x Position	0.04	0.04	0.03	0.04	0.00	0.04	0.00	0.04
Watson Glazer (WG)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
WG x Position	-0.01	0.01	-0.01	0.01	0.00	0.01	0.00	0.01
Wesman Quantitative (WQ)	0.01	0.01	0.01	0.01	0.00	0.01	0.00	0.01
WQ x Position	0.02	0.02	0.02	0.02	0.01	0.02	0.01	0.02
Wesman Verbal (WV)	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
WV x Position	0.01	0.02	0.00	0.02	0.00	0.02	0.00	0.02
General impression (GI)			0.37	0.13 **			0.33	0.12 **
GI x Position			0.44	0.28			0.04	0.27
Preparation (P)			0.05	0.05			0.02	0.04
P x Position			-0.08	0.10			0.04	0.10
Overall recommendation (OR)			-0.14	0.09			-0.04	0.09
OR x Position			-0.11	0.20			0.01	0.20
R-square	0.04		0.07		0.05		0.07	

Note. Model 1 *df* = 479. Model 2 *df* = 473. *p < .05. **p < .01. *N* = 505.

relationship with performance for police officers (β = -.10). When assessor judgments were added to the model, they accounted for an additional 3% of variance in work disposition. Only the assessor rating of general impression showed a significant relationship with performance (β = .16).

The overall level of prediction was similar for the second performance dimension (social/emotional competence). The battery of tests accounted for 5% of the variance in the performance. Extroversion was significantly and positively related to performance across both positions (β = .11). The Firo Control-Wanted scale was significantly and negatively related to performance across both positions (β = -.12). Conscientiousness significantly interacted with position, showing a positive relationship with performance among firefighters (β = .13) and a negative relationship among police officers (β = -.08). When assessor judgments were added to the model, the accounted for an additional 2% of variance in social/emotional competence. Only the assessor general impression rating was significantly related to performance (β = .15).

To address the third research question, the full model was reestimated allowing the slopes for all predictors to vary across municipalities and allowing the slopes for the assessor ratings to vary across assessors. For each predictor, this analysis provided an estimate of the average slope and the *SD* of slopes across municipalities or assessors. Because the research question involves assessor differences, we focus here only on the random assessor effects. The full results for this analysis are presented in the Appendix.

None of the assessor variance components were significant for the first performance dimension. For the prediction of the social/emotional competence, we found significant differences across assessors in the slopes for general impression ratings, slope $SD = 0.27$, $\chi^2(4) = 10.47$, $p = .03$. To further explore this result, we computed empirical Bayes estimates of the general impression slope separately for each assessor. Although on average the standardized slope was .16, the assessor-specific slopes varied considerably. Whereas the general impression rating from some assessors demonstrated a useful level of incremental validity over the test battery (β = .25), for others incremental validity was near zero (β = .03).

Although the limited number of assessors prevented any systematic analysis of assessor differences, we conducted a post-hoc examination of assessor specific slopes for patterns related to assessor background. The six assessors who performed had a range of experience conducting IPA (from 2 years to over 25 years), and different educational backgrounds (MS, PhD, or PsyD in industrial-organizational psychology, clinical psychology, or counseling psychology). The two individuals with the most experience, who both had PhDs, showed greater incremental validity than most of the other assessors. At the same time, one of the least experienced assessors with a master's degree also performed at a similar level and outperformed some with a PhD and substantially more experience. No field of study consistently showed higher slopes than others. Overall, no clear pattern of assessor differences due to experience or education could be discerned.

Effects of Statistical Artifacts

Our estimates of criterion-related validity involved no correction for statistical artifacts such as criterion unreliability or range restriction. The criterion measure was developed for the current study, and the reliability of this measure is unknown. If one were to apply a correction based on typical values reported in the literature (e.g., criterion reliability of .52; Viswesvaran, Ones, & Schmidt, 1996), the correlations with job performance would be about 39% larger.

Similarly, range restriction may have occurred both before and after the assessments were conducted, further reducing the observed validities. Prior to the IPA, candidates were prescreened by the municipalities, and candidates were hired in part based on the results of the IPA.

Although we could not assess the degree of range restriction due to pre-screening, we were able to examine range restriction that occurred after the IPA by comparing the full sample to those where were ultimately hired. In the current data, substantial range restriction was observed on the assessor general impression rating and overall recommendation, where the *SDs* among those hired was 68% and 65% of the full sample, respectively. The degree of range restriction on all other variables was trivial, with restricted *SD* > 94% of the unrestricted *SD*.

Correction for direct range restriction somewhat increased the size of bivariate correlations. The general impression rating correlated .13 with both dimensions of performance, and correction for direct range restriction increased these to .19. Similarly, correction for range restriction increased the validity of the overall recommendation from .05 to .08 for work disposition, and from .08 to .12 for social/emotional competence. In all cases, correction for range restriction did not substantially change the size of the validity coefficient.

DISCUSSION

IPA represents an attractive approach to employee selection, utilizing the experience of an expert psychologist to interpret test scores, conduct an extensive interview, and integrate the multiple pieces of information into a coherent picture of the candidate. Despite its appeal, the literature provides only limited empirical evidence in support of IPA. Questions remain regarding how assessors utilize IPA data and how well assessor recommendations predict employee job performance. The current study sought to address these questions through an extensive analysis on one firm's assessment practices for public safety positions.

First, we examined the drivers of assessor judgments, that is, which components are given the most weight in assessor ratings. To our surprise, assessor ratings were only weakly related to the test results. The combination of seven personality scores and three ability tests accounted for only 3% to 9% of the variance in assessor dimension ratings, and 6% of the variance in overall recommendations.

Our ability to model how assessors form judgments was limited by the lack of full data on the assessment components. We had access only to candidate scores on standardized tests. In addition to these test scores, assessors had access to two additional sources of information, the interview and biodata form, which we were unable to include in our model. Although this substantially limits what we can conclude about the assessor judgment process, it is still noteworthy that ratings were not strongly driven by personality traits or general cognitive ability.

Another possible explanation for the weak prediction of assessor ratings is that the linear regression model did not adequately represent the way assessors use the test information. Assessors may recognize configurations of test scores that are more complex than the linear additive process represented by the regression model (Highhouse, 2002). However, configural rules generally do not play much of a role in decision making (Karren & Barringer, 2002). Therefore, it is questionable whether modeling more complex decision rules would have substantially changed our results.

Personality scales had the most influence on the assessment dimension ratings. Specifically, the Neuroticism, Agreeableness, and Conscientiousness scales were significant predictors of the first assessment dimension. Assessor judgments were not related to the cognitive ability tests, despite the substantial evidence for the validity of general cognitive ability in public safety (Aamodt, 2004; Barrett et al., 1999; Henderson, 2010).

Regarding criterion-related validity, the results were mixed. Assessor dimension ratings comprising the general impression dimension were found to add significantly to the prediction of both dimensions of job performance. However, assessor ratings of job preparation and their overall recommendations were not related to performance measures. It is interesting to note that while general impression and overall recommendation were highly correlated, general impression, which was a mechanical combination of expert ratings, was more predictive than the subjective integration of this information as reflected in overall recommendation. This is consistent with the literature on clinical versus statistical prediction, which has generally found that expert judgment can be useful in assessing specific competencies, but that when it comes to integrating information, mechanical combination tends to outperform expert judgment (Holt, 1958; Huse, 1962; Meyer, 1956; Trankell, 1959). Holt (1958) found that using a combined technique of clinical and mechanical prediction was more effective than making a purely clinical prediction about a candidate. Further research is recommended to identify the optimal combination of expert judgment and mechanical combination in assessment practice.

The incremental validity of assessor ratings was small, accounting for 2% to 3% of the variance in supervisor ratings of performance. The magnitude of the standardized coefficients for General Impression (.15–.16) was consistent with meta-analytic estimates of the IPA validity for nonmanagerial jobs (uncorrected r = .19; Morris et al, 2015). Although modest, the finding of incremental validity stands in contrast to past research that has failed to support incremental validity of assessors over the test battery (Holt, 1958; Huse, 1962; Meyer, 1956; Trankell, 1959).

Our ability to assess incremental validity was limited by the lack of data on the interview and biodata tools, which therefore could not be included as separate predictors in the regressions models. It may be that the validity gains were due to the inclusion of these predictors. Still, the results suggest that the validity of IPA cannot be attributed solely to cognitive and personality traits that can be readily assessed via standardized tests.

Our results also highlight the importance of examining differences in IPA validity across assessors. Although the average regression coefficients were weak, the magnitude of the regression coefficients for assessor general impression ratings predicting social/emotional competence showed significant variability across assessors. Thus, assessor ratings had moderately strong incremental validity for some of the assessors (with a standardized slope as large as .25), but the relationship was essentially zero for other assessors.

The question of individual differences in validity has previously been raised in the context of structured interviews, but the research in this area has not supported differences in validity across interviewers (Pulakos, Schmitt, Whitney, & Smith, 1996; Van Iddekinge, Sager, Burnfield, & Heffner, 2006). It may be that greater complexity of the IPA process (i.e., requiring the integration of test scores, interview response, etc.), combined with the relatively high degree of discretion allowed assessors in the IPA context, increases the impact of individual differences in assessor skill and amplifies differences in assessor validity.

The six assessors who performed the IPAs in the current study reflected diverse backgrounds and training. They had a range of experience conducting IPA (from 2 years to over 25 years), and different educational backgrounds (MS, PhD, or PsyD in industrial-organizational psychology, clinical psychology, or counseling psychology). Although our data did not contain enough assessors to test systematic differences across backgrounds, it has been theorized that these differences in assessors could impact the validity of the IPA (Ryan & Sackett, 1992). We hope that future research examining a larger collection of assessors will be able to identify the types of training and experience that are associated with higher assessor validity.

Limitations

Although public safety selection is an important context in which IPA is used, our results may not generalize to other types of occupations. In particular, IPA is widely used in executive selection (Silzer & Jeanneret, 2011), and IPA may provide higher validities for managerial occupations (Morris et al, 2015). Therefore, more favorable validity evidence might be expected for positions with greater managerial responsibility.

It is possible that the weak prediction of performance in this sample is partly due to limitations of the criterion measure. Researchers have noted the difficulty of collecting accurate performance data in public safety occupations, because supervisors in this context often have limited opportunity to observe their subordinates' performance (Hirsh, Northrop, & Schmidt, 1986).

The sample of public safety officers whose performance was evaluated was small within each municipality in relation to the number of predictors in the analysis, which presented a limitation in the data analyses. Moreover, the number of assessors and municipalities was also smaller than ideal for multilevel analysis, particularly when testing variance components (Hox, 2002). Furthermore, potential differences that may exist across assessors could not be tested due to the small number of assessors.

CONCLUSION

The results of this study suggest there is some utility to the IPA process. Assessor judgments were to add to the prediction of job performance over the test battery. At the same time, the strength of the predictive relationships was fairly low, and assessments accounted for only a small proportion of the variance in job performance.

Concerns about the utility of assessments are heightened by the finding that assessor judgments were largely unrelated to the test results. Much of the appeal of IPA is that the judgments are based on a comprehensive assessment of the candidate (Highhouse, 2002). Considerable time and cost are devoted to the administration of the test battery, and one must question the utility of this extensive assessment process if it has so little effect on the final evaluation. We are not, of course, recommending that the use of standardized tests be eliminated (in effect turning the assessment into an interview). On the contrary, we would argue that the validity of assessments can be enhanced by ensuring more consistent and extensive use of test results in the assessor ratings and recommendations (McPhail & Jeanneret, 2012).

REFERENCES

Aamodt, M. G. (2004). *Research in law enforcement selection.* Boca Raton, FL: Brown Walker.

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S....Rush, J.D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34,* 341-382.

Barrett, G. V., Polomsky, M. D., & McDaniel, M. A. (1999). Selection tests for firefighters: A comprehensive review and meta-analysis. *Journal of Business and Psychology, 13*(4), 507-513.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Camerer, C. F., & Johnson, E. J. (1991). The process-performance paradox in expert judgment: How can experts know so much and predict so badly? In K. Anders-Ericsson, and J. Smith, (Eds.), *Toward a general theory of expertise* (pp. 195-217). New York, NY: Cambridge University Press.

Consulting Psychologists Press. (1989). *Firo-B.* Palo Alto, CA: Author.

Costa, P. T., Jr. & McCrae, R. R. (1991). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional Manual.* Odessa, FL: Psychological Assessment Resources, Inc.

Garb, H. (2003). Clinical judgment and mechanical prediction. In J. R. Graham, & J. A. Naglierie (Eds.), *Assessment psychology* (pp 27-42). Hoboken, NJ: John Wiley& Sons Inc.

Grove, W., Zald, D.H., Lebow, B.E., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19-30.

Hammer, A. L., & Schnell, E. R. (2000). *FIRO-B® technical guide.* Mountain View, CA: CPP, Inc.

Henderson, N. D. (2010). Predicting long-term firefighter performance from cognitive and physical ability measures. *Personnel Psychology, 63*, 999-1039.

Highhouse, S. (2002). Assessing the candidate as a whole: A historical and critical analysis of individual psychological assessment for personnel decision making. *Personnel Psychology, 55*, 363-396.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 333–342.

Hilton, A., Bolin, S. F., Parker, J. W., Taylor, E. L., & Walker, W. B. (1955). The validity of personnel assessments by professional psychologists. *Journal of Applied Psychology, 39*, 287-293.

Hirsh, H. R., Northrop, L. C., & Schmidt, F. L. (1986). Validity generalization results for law enforcement occupations. *Personnel Psychology, 39*(2), 399-420.

Holt, R. (1958). Clinical and statistical prediction: A reformulation and some new data. *Journal of Abnormal and Social Psychology, 56*, 1-12.

Hox, J. (2002). *Multilevel Analysis Techniques and Applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Huse, E. (1962). Assessments of higher-level personnel: IV. The validity of assessment techniques based on systematically varied information. *Personnel Psychology, 15*, 195-205.

Jeaneret, R., & Silzer, R. (1998). An overview of individual psychological assessment. In R. Jeanneret & R. Silzer (Eds.), *Individual psychological assessment: Predicting behavior in organizational setting* (pp. 3-26). San Francisco, CA: Jossey-Bass.

Karren, R. J., & Barringer, M. W. (2002). A review and analysis of policy-capturing methodology in organizational research: Guidelines for research and practice. *Organizational Research Methods, 5*, 337-361.

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology, 98*(6), 1060-1072.

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained variance measures for multilevel models. *Organizational Research Methods, 17*, 433-451.

McPhail, S. M., Jeanneret, P. R. (2012). Individual psychological assessment. In N. Schmitt (ed.), *Oxford handbook of personnel, assessment, and selection* (pp. 411-442). New York, NY: Oxford University Press:

Meyer, H. (1956). An evaluation of a supervisory selection program. *Personnel Psychology, 9*, 499-513.

Miner, J. (1970). Psychological evaluations as predictors of consulting success. *Personnel Psychology, 23*, 393-405.

Morris, S. B., Daisley, R. R., Wheeler, M., & Boyer, P. (2015). A meta-analysis of the relationship between individual assessments and job performance. *Journal of Applied Psychology, 99* (3). Advance online publication. http://dx.doi.org/10.1037/a0036938.

Morris, S. B., Kwaske, I. H., & Daisely, R. R. (2011). Validity of individual psychological assessments. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 4*, 322-326.

O'Brien, J, & Rothstein, M. G., (2011). Leniency: Hidden threat to large-scale, interview-based selection systems. *Military Psychology, 23*, 601-615.

Prien, E.P., Schippmann, J. S., & Prien, K. O. (2003). *Individual assessment as practiced industry and consulting*. Mahwah, NJ: Erlbaum.

Psychological Corporation. (1980). *Watson-Glaser critical thinking apprais*al. Orlando, FL: Harcourt Brace.

Pulakos, E. D., Borman, W. C., & Hough, L. M. (1988). Test validation for scientific understanding: Two demonstrations of an approach to studying predictor-criterion linkages. *Personnel Psychology, 41*, 703-716.

Pulakos, E. D., Schmitt, N., Whitney, D., & Smith, M. (1996). Individual differences in interviewer ratings: The impact of standardization, consensus discussion, and sampling error on the validity of a structured interview. *Personnel Psychology, 49*, 85-102.

Raudenbush, S. W., Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2004). *HLM6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International, Inc.

Ryan, A. M., & Sackett, P. R. (1987). Exploratory study of individual assessment practices: Interrater reliability and judgments of assessor effectiveness. *Journal of Applied Psychology, 74*, 568-579.

Ryan, A. M., & Sackett, P. R. (1992). Relationships between graduate training, professional affiliation, and individual psychological assessment practices for personnel decisions. *Personnel Psychology, 45*, 363-387.

Ryan, A. M., & Sackett, P. R. (1998). Individual assessment: The research base. In R. Jeanneret & R. Silzer (Eds.), *Individual psychological assessment: Predicting behavior in organizational settings* (pp. 54 – 87). San Francisco, CA: Jossey-Bass.

Schmidt, F., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bullentin, 2*, 262-274.

Silzer, R. J., & Jeanneret, R. (2011). Individual psychological assessment: A practice and science in search of common ground. *Industrial and Organizational Psychology: Perspectives on Science and Practice,4*, 270-296.

Trankell, A. (1959). The psychologist as an instrument of prediction. *Journal of Applied Psychology, 43*, 170-175.

Van Iddekinge, C. H., Sager, C. E., Burnfield, J. L., & Heffner, T. S. (2006). The variability of criterion-related validity estimates among interviewers and interview panels. International *Journal of Selection and Assessment, 14*, 193-205.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557-574.

Weiner, I. B. (2003). The assessment process. In J. R. Graham & J. A. Naglierie (Eds.), *Assessment psychology* (pp.3-25). Hoboken, NJ: John Wiley & Sons Inc.

Wesman, A. G. (1965). *Wesman personnel classification test manual*. New York: The Psychological Corporation.

RECEIVED 10/21/14 **ACCEPTED** 9/23/15

APPENDIX A

TABLE A1.
Descriptive Statistics

	Police			Fire		
	N	*M*	*SD*	*N*	*M*	*SD*
Neuroticism	1189	2.07	.78	476	2.18	.75
Extraversion	1189	3.61	.71	476	3.83	.75
Openness	1189	2.99	.82	476	3.11	.84
Agreeableness	1189	3.19	.77	476	3.70	.78
Conscientiousness	1189	3.73	.79	476	3.65	.83
Firo Control-Expressed	1391	2.73	2.36	555	2.07	2.00
Firo Control-Wanted	1391	1.92	1.52	555	2.78	2.07
Watson Glaser	1391	55.45	10.66	553	57.73	10.05
Wesman Quantitative	1389	8.29	3.72	554	9.83	3.80
Wesman Verbal	1390	22.42	5.90	553	24.39	5.63
General impression	1410	2.94	.48	563	2.98	.47
Preparation	1410	2.97	.73	565	3.04	.86
Overall recommendation	1410	2.77	.64	567	2.83	.63
Selection decision	1410	.49	.50	562	.46	.50
Performance: work Disposition	435	3.81	.79	173	3.83	.69
Performance: social/emotional competence	435	3.79	.76	173	3.64	.64

TABLE A2.
Fixed Effect Estimates from the Hierarchical Linear Model Predicting Performance Ratings from Assessment Components and Assessor Ratings with Random Slopes

	Work Disposition		Social Emotional Competence	
Predictor	*Coeff.*	*SE*	*Coeff.*	*SE*
Intercept	3.83	0.06 **	3.76	0.06
Intercept x Position	-0.02	0.13	0.11	0.13
Neuroticism	0.11	0.05 *	0.10	0.05
Neuroticism x Position	-0.06	0.12	0.11	0.12
Extraversion	0.10	0.05	0.12	0.06 *
Extraversion x Position	0.01	0.11	0.12	0.13
Openness	-0.08	0.04	-0.08	0.04 *
Openness x Position	0.03	0.09	0.05	0.09
Agreeableness	0.01	0.05	-0.04	0.05
Agreeableness x Position	0.22	0.12	0.23	0.11 *
Conscientiousness	-0.04	0.05	-0.01	0.04
Conscientiousness x Position	-0.30	0.12 *	-0.28	0.10 *
Firo Control-Expressed	0.00	0.02	0.00	0.02
Firo Control-Expressed x Position	0.08	0.04 *	0.05	0.04
Firo Control-Wanted	-0.02	0.02	-0.03	0.02
Firo Control-Wanted x Position	0.04	0.04	0.03	0.04
Watson Glaser	0.00	0.00	0.00	0.00
Watson Glaser x Position	-0.01	0.01	0.00	0.01
Wesman Quantitative	0.01	0.01	0.00	0.01
Wesman Quantitative x Position	0.02	0.03	0.01	0.03
Wesman Verbal	0.01	0.01	0.01	0.01
Wesman Verbal x Position	0.00	0.02	-0.01	0.02
General impression	0.46	0.17 *	0.41	0.21 *
General impression x Position	0.37	0.32	-0.02	0.36
Preparation	0.04	0.05	0.00	0.06
Preparation x Position	0.03	0.11	0.13	0.11
Overall recommendation	-0.13	0.13	-0.03	0.13
Overall recommendation x Position	-0.23	0.25	-0.09	0.24

Note. Random slopes were estimated for all predictors across municipalities and for assessor ratings across assessors. *df* = 473. **p* < .05. ***p* < .01.

TABLE A3.

Random effect estimates from hierarchical linear model predicting performance ratings from assessment components and assessor ratings

	Work Disposition		Social/Emotional Competence	
Random Effect	*SD*	*Chi-sq*	*SD*	*Chi-sq*
Municipality Effects				
Intercept	0.28	49.67 **	0.28	62.87 **
Neuroticism	0.18	36.79 **	0.20	30.59 **
Extraversion	0.13	26.05 *	0.24	29.83 **
Openness	0.12	18.80	0.11	23.05
Agreeableness	0.16	27.86 *	0.17	25.85 *
Conscientiousness	0.18	18.13	0.09	15.92
Firo Control-Expressed	0.05	26.00 *	0.06	30.33 **
Firo Control-Wanted	0.05	21.99	0.06	22.38
Watson Glaser	0.01	21.65	0.02	25.50 *
Wesman Quantitative	0.04	25.25 *	0.05	18.43
Wesman Verbal	0.03	27.47 *	0.03	41.18 **
General Impression	0.49	37.64 **	0.72	44.49 **
Preparation	0.17	43.16 **	0.19	45.59 **
Overall Recommendation	0.43	46.57 **	0.46	61.99 **
Assessor Effects				
Intercept	0.03	3.51	0.04	5.83
General Impression	0.18	4.81	0.27	10.47 *
Preparation	0.02	2.41	0.05	3.32
Overall Recommendation	0.13	5.83	0.17	8.29

Note. For municipality effects, *df* = 14. For assessor effects, *df* = 4. **p* < .05. ***p*<.01.