Personality, Biodata, and Situational Judgment Tests: A Summary of the Research Literature Comparing Predictive vs. Concurrent Validity

Report Prepared for:
Jeff Weekley
Kenexa
11/19/01

Prepared by:
Crystal Harold
Robert E. Ployhart
Department of Psychology
George Mason University
Fairfax, VA 22030

Overview:

Research has indicated that Personality, Biodata, and Situational Judgment Tests have all been found to predict job performance (usually with less adverse impact than cognitive ability tests). However, it is important to differentiate between studies using applicant samples and incumbent samples, due to the possibility that responses to these tests may differ according to these sample (e.g., different levels of motivation between applicants and incumbents). Included below are findings reported from various articles as to differences in predictive and concurrent validities of the selection measures of interest. These articles indicate:

- Personality: Predictive validities are slightly lower than concurrent validities (Hough, 1998). Other researchers contend that predictive and concurrent validities for personality-based integrity tests are fairly similar (Ones, Viswesvaran, Schmidt 1993).

- Biodata: Most biodata research to date has focused on "faking" and differences in biodata scores between incumbents (instructed to fake or respond honestly) and applicants. Researchers have indicated that applicants (who were not hired for the given position) and incumbents (instructed to respond honestly) obtain similar mean scores on a given biodata measure (Becker and Colquitt, 1992). However, many note that research comparing predictive and concurrent validities of biodata measures is needed.

- Situational Judgment Tests: A meta-analysis of situational judgment tests found predictive validity was lower than concurrent validity. These findings, however, may be due to sample size artifacts (six predictive validity designs compared to 96 concurrent validity designs, respectively; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Further research in this area is warranted.

In the following pages we review specific articles comparing predictive and concurrent validities for personality, biodata, and situational judgment tests. These references are presented in chronological order.

Personality

Ellingson, J.; Smith, B.; and Sacket, P. (2001). Investigating the influence of social desirability on personality factor structure. Journal of Applied Psychology, 86, 122-133.

This article investigated whether socially desirable responding effects the factor structure of personality measures. Respondents to these measures were placed in one of two groups. One group consisted of individuals identified as responding honestly, the other group consisted of individuals identified as responding in a socially desirable manner. Results indicated that the factor structures did not differ between the two groups, and thus that social desirability did not influence the factor structures of the personality measures.

Robie, C.; Zickar, M.; and Schmit, M. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. Human Performance, 14, 187-207.

Robie and colleagues examined the differences in responding to 6 personality scales, between incumbents and applicants. In addition the authors applied an item response theory model to detect aberrant responding. Results indicated that applicants obtained significantly higher mean scores than incumbents on all 6 scales (about .5 standard deviation higher). However, only one of the six scales indicated items functioned differently between the two groups (Work Focus Scale), and none of the scales functioned differently across the groups. These results suggest that while scores between applicants and incumbents are significantly different, these differences are not due to aberrant responding.

Stark, S.; Chernyshenko; Chan, K.; Lee, W.; and Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. Journal of Applied Psychology, 86, 943-953.

Stark and colleagues administered the Sixteen Personality Factor Questionnaire (16PF) to applicants and nonapplicants, to examine the effects of faking on test scores. Results indicated that scale reliabilities were slightly lower for applicants than nonapplicants. Applicants scored higher on personality scales (such as conscientiousness and agreeableness). In addition the authors used different item response theory models (IRT) to detect faking. Their results indicated that differential item/test functioning occurred across applicant and nonapplicants. This suggests that faking affects the construct validity of personality scales and also that research that compares faking by creating groups from a single sample using their impression management (IM) scores may not be generalizable.

Smith, B.; Hanges, P.; & Dickson, M. (2000). Personnel selection and the five-factor model: Reexamining the effects of applicants frame of reference. Journal of Applied Psychology, 86, 2, 304-315.

The authors address criticisms about the adequacy of the five-factor model in describing job applicant personality. Using a student sample, applicant sample, and an incumbent sample, the authors found that the five-factor model did fit all samples well (though it fit applicants slightly better than the student sample). Citing results found by Barrett et al (1981) (who found no difference in predictive and concurrent validities for cognitive ability tests) the authors suggest that there are also no differences between these validation designs on the five-factor model. (Note, applicants and incumbents came only from positions Holland's Realistic and Conventional classifications. The authors caution that their results may not hold across occupations).

Hurtz, G. and Donovan, J. (2000). Personality and job performance: The big five revisited. Journal of Applied Psychology, 85, 869-879.

The authors suggest that previous meta-analyses examining the predictive validity of Big 5 personality dimensions contain threats to construct validity, citing inclusion of data not derived from Big 5 measures. Results yielded validities similar to those obtained by Barrick and Mount (1991), though validity coefficients for conscientiousness were slightly lower in this study (about .20). Despite this, they suggest that personality variables might not contribute greatly to selection procedures (stipulating that other techniques, such as interviews, already captured aspects of personality). In addition concerns of faking, and negative applicant reactions, may add to the argument that the Big 5 provides a small contribution to the selection process.

Rosse, J.G.; Stecher, M.D.; Miller, J.L, & Levin, R.A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. Journal of Applied Psychology, 83, (4), 634-644.

Rosse and colleagues examined whether applicants are likely to practice impression management and the implications of faking on hiring decisions. The authors administered a personality inventory to applicants as part of the application process. The same inventory was administered to incumbents (in the same organization) of the given positions. Results indicated that applicants did practice response distortion and, on average, obtained scores at least one standard deviation above the scores of incumbents. Also of interest was that there were differences in the degree to which applicants practiced response distortion, with some of the applicants scoring two to three standard deviations above the mean.

Hough, L.M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. Human Performance, 11, 209-244.

Hough discusses self-description inventories and their validity in predicting job performance. She indicates that when directed faking is present, criterion-validity is near zero, and also that both concurrent and predictive designs yield criterion-validity. She examined the effects of faking on criterion-validity of self-description measures, and offered two strategies to remedy the effects of response distortion. The first strategy involved correcting content scores based on scores on the Unlikely Virtues (UV) scale. The second strategy entailed removing individuals, based on their UV scores, from the applicant pool. Hough reported that implementing these strategies caused more highly similar applicant and incumbent mean scores (content scores), no adverse impact, and no affect on criterion-related validity.

Hough, L. M. (1998). Personality at work: Issues and evidence. In M. D. Hakel (Ed.), Beyond Multiple Choice: Evaluating Alternatives to Traditional Testing for Selection (pp.131-166). Mahwah, NJ: Lawrence Erlbaum Associates.

Hough discussed her work on Project A, for which she collected both incumbent and applicant data. The criterion measures for which she collected data included job proficiency, educational success, counterproductive behavior, and training success. (Note, Hough did not use the Big Five dimensions as her measure of personality. She included 8 dimensions; affiliation, potency, achievement, dependability, adjustment, agreeableness, intellectance, and rugged individualism. Her research has shown that achievement (thought to be a subset of conscientiousness) is the best predictor of all criteria of interest).

Comparing concurrent to predictive validities, she found that the concurrent validity coefficients for Achievement were .13, .35, and -.42 for job proficiency, educational success, and counterproductive behaviors, respectively. Predictive validity coefficients for Achievement were .19, .19, .23, and -.33 for job proficiency, training success, educational success, and counterproductive behaviors, respectively. In addition, Achievement, Dependability, and Adjustment correlated -.42, -.39, and -.39 with Counterproductive behaviors for the concurrent validity samples, but. -33, -.23, and -.17 in the predictive validity samples. Also, Educational Success correlated .35 and .32 with Achievement and Adjustment in the concurrent samples, and .23 and .21 in the predictive samples.

The above correlations indicate that validities between concurrent and predictive validity samples are different, with predictive validity coefficients usually (but not always) being lower than concurrent coefficients. Hough estimated that "concurrent validity studies produce validity coefficients that are, on average, .07 points higher than predictive validity studies." When these comparisons are done within criterion construct, the mean differences range from .04 to .15.

Please see the tables on the following page for a comparison.

Predictive Validities

Personality Constructs	Job Proficiency	Training Success	Educational Success	Counterproductive Behavior
Potency	.07	.07	.15	-.04
Achievement	.19	.19	.23	-.33
Dependability	.04	.11	.13	-.23
Adjustment	.05	.11	.21	-.17
Agreeableness	.01	.10	.00	-.01
Intellectance	-.02	.02	.12	.24
Rugged Individualism	-.05	.02	-.03	.00

Concurrent Validities

Personality Constructs	Job Proficiency	Training Success	Educational Success	Counterproductive Behavior
Potency	.10	.19	.19	-.24
Achievement	.13	**----**	.35	-.42
Dependability	.09	.12	.28	-.39
Adjustment	.11	.03	.32	**-.**39
Agreeableness	.08	-.05	.02	-.21
Intellectance	.05	.20	.26	**----**
Rugged Individualism	.13	.18	-.18	.02

Hough, L. and Ones, D. (in press). *The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In N. Anderson, D.S. Ones, K. Sinangil, & C. Viswesvaran (Eds.), International handbook of work and organizational psychology, Sage Publications*

These authors also discuss the predictive and concurrent validity of personality variables. Namely, the authors respond to criticisms that personality tests are fakable, which affects the validity of these measures as selections tools. Using a meta-analysis conducted by Ones, Viswesvaran, and Schmidt (1993), the authors suggest that personality measures are valid (despite the possibility of faking) and cite predictive validity coefficients of .29 for applicant studies, and .26 for studies involving incumbents. In addition the concurrent validity for studies involving incumbents is .29. (These are the coefficients for predicting counterproductive behavior).

Ones, D.S.; Viswesvaran, C., & Reiss, A.D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. Journal of Applied Psychology, 81, (6), 660-679.

Among the goals of Ones, Viswesvaran, and Reiss (1996) was to determine whether social desirability (impression management) is a cause for concern to industrial-organizational psychologists. The authors were also interested in ascertaining whether social desirability was a factor representing actual differences in conscientiousness and emotional stability between individuals. Lastly the authors examined whether social desirability functioned as a predictor, moderator, or suppressor variable for job performance. In their meta-analyses of the literature, the authors found that social desirability correlated with personality dimensions, suggesting that scores on social desirability scales reflect actual differences in personality.

In addition, it was found that ability to respond to items in a socially desirable manner relates to emotional stability and conscientiousness. Therefore, individuals who obtained higher scores on the emotional stability and conscientiousness scales (two seemingly positive characteristics) also exhibited higher levels of socially desirable responding. Such a finding suggests that socially desirable responding may not be a detrimental practice. The authors also found that social desirability did not function as a predictor, moderator, or suppressor variable for the job performance criterion, again suggesting that socially desirable responding on personality measures does not hurt the validity of such measures.

Schmit, M. and Ryan, A.M. (1993). The big five in personnel selection: Factor structure in applicant and nonapplicant population. Journal of Applied Psychology, 78, 966-974.

The authors investigated the 5-factor model of personality using both applicant and nonapplicant samples. Results of confirmatory factor analyses indicated that 5-factor model structure did not fit the applicant sample (but did fit the student nonapplicant sample). For the applicant sample, an additional dimension emerged (the ideal-employee

factor). In addition, the authors suggest that the subscales of the NEO-FFI may not be best for employee selection.

Ones, D.; Viswesvaran, C.; and Schmidt, F. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. Journal of Applied Psychology, 78, 679-703.

Results of a meta-analysis suggest that integrity test validities predict job performance and counterproductive behaviors. Personality measures correlated with supervisory ratings of job performance and externally and self-reported counterproductive behaviors. The authors also conclude that integrity tests have generalizable validity. However, a comparison of predictive versus concurrent designs using applicants and employees, yielded different validities between the groups (for predicting supervisory ratings of job performance). The true validity coefficients for predictive designs were .41 and .26 for applicant and incumbent samples, respectively. For concurrent designs the true validity coefficients were .48 and .37 for applicants and incumbents, respectively.

	Applicant	**Employees**
	Predictive	
Mean r	.25	.15
True Validity (ρ)	.41	.26
	Concurrent	
Mean r	.29	.22
True Validity (ρ)	.48	.37

Barrick, M. and Mount, M. (1991). The big five personality dimensions and job performance: A meta-analysis. Personnel Psychology, 44, 1-25.

This meta-analysis that examined the relationship of the Big 5 personality dimensions to job performance criteria across a number of jobs. Results indicated that conscientiousness was a valid predictor across all criteria and jobs. Extraversion and openness to experience were valid predictors of training proficiency. These results provide support for the use of the Big 5 in selection, and also that it is both a be valid predictor of performance and generalizable across different occupations.

Biodata

Carlson, K.; Scullen, S.; Schmidt, F.; Rothstein, H.; and Erwin, F. (1999). Generalizable biographical data validity without multi-organizational development and keying. Personnel Psychology, 52, 731-753.

The authors found that the biodata measure, the Manager Profile Record (MPR), was a valid predictor of promotional progress (of managers), and generalized across 24 organizations. It was also a valid predictor across ages, genders, and educational levels. The estimated true validities for the groups were: .53 across organizations, .50 across genders, .49 across education levels, and .51 across ages. The authors suggest that well-developed biodata items are generalizable and can be used across organizations.

Mumford, M. and Whetzel, D. (1997). Background Data. In D. Whetzel and G. Wheaton (eds.), Applied Measurement Methods in Industrial Psychology, Davies-Black Publishing.

Biodata measures are considered effective predictors of job performance (the criterion-related validity coefficients are usually in the .40-.50 range). The authors suggest that while faking can be a problem on biodata items, including "impossible life event" items can help detect fakers (i.e. winning an award that doesn't exist). In addition, research suggests that faking will occur more when the applicants have a clear idea of what the job entailed, thus enabling them to answer, as they believed an ideal candidate would. Mumford (1994) indicated that faking occurred less when applicants couldn't find a clear right or wrong answer. The authors also contend that biodata measures have less adverse impact than do other selection measures. They cite (Reilly and Chao, 1982), who after reviewing 11 background data studies found that there were no significant differences among the ethnic subgroups.

Mumford, M. and Stokes, G. (1990). Developmental determinants of individual action: Theory and practice in applying background measures. In M. Dunnette and L. Hough (eds.) Handbook of Industrial and Organizational Psychology, Palo Alto, Consulting Psychologists Press.

The chapter reviews the history of biodata measures. In general, most studies on biodata have indicated that have good predictive validity (about .30-.35) of a wide range of criteria (such as job performance, satisfaction, turnover). Citing Barge and Hough (1986), the authors reveal that validity coefficients were higher for concurrent studies than they were for predictive studies. Mumford and Stokes also report that biodata measures are among the best predictors for training and job performance, with validity coefficients ranging between .25 and .37. In addition, biodata measures yield validity coefficients that show very little difference between different ethnicities.

Schmidt, F. and Rothstein, H. (1994) Application of Validity Generalization to Biodata Scales in Employment Selection. In G. Stokes, M. Mumford, & W. Owens (eds.), Biodata Handbook, Consulting Psychologists Press, Inc.

This chapter gives a brief overview of the history of biodata research. The authors conclude that biodata scales that are generalizable across organization can be constructed. (Previous research suggested that biodata might only be valid if it's created for specific jobs within an organization and thus not generalizable). The authors also discuss job experience as it relates to biodata validity. Other researchers have maintained that biodata validity in concurrent studies is the result of measuring knowledge gained through job experience. If this were the case than concurrent validities would be higher than predictive validities. However, the authors suggest that if job experience is held constant, this may not occur. The authors discuss a meta-analysis, in which job experience was held "nearly" constant. The findings indicated that mean validities did not decline for ability or duty ratings, suggesting that biodata validity may not be an artifact of job experience. The authors caution, however, "individuals differences in job experience is not the same as the question of whether concurrent validities are the same as predictive validities…" This chapter doesn't discuss any known articles comparing predictive and concurrent biodata validities and conclude such research "would be quite useful."

In addition the authors found that biodata measures did not have to be situation specific, and generalized across job. In a meta-analysis using 79 validity coefficients, they found a mean true validity of .36 and .34 for ability for perform and performance of duties, respectively.

Stokes, G.; Hogan, J.; and Snell, A. (1993). Comparability of incumbent and applicant samples for the development of biodata keys: The influence of social desirability. Personnel Psychology, 46, 739-762.

Both applicants and incumbents completed a biodata measure. Raters were then asked to rate responses on this measure according to how socially desirable each response option was. Results indicated that both applicants and incumbents practiced socially desirable responding, but applicants to a greater extent. Most important for our purposes, was that the authors found that there were no commonly scored items between the two samples, indicating that biodata created for incumbents may not generalize to applicants.

Becker, T. and Colquitt, A. (1992). Potential versus actual faking of a biodata from: An analysis along several dimensional item type. Personnel Psychology, 45, 389-406.

This study examined faking on a biodata questionnaire. Participants included applicants and incumbents. Incumbents were instructed to take the test, half of which were told to fake ("fake-take") and the other half to respond honestly ("straight take"). Applicants completed the measure as if it were just part of the applicant process ("real

take", the group was later broken down into hires—those who received a job offer and applicants---those who didn't).

The mean scores on the biodata questionnaire were interesting for the different conditions. The mean scores were, 17.71, 23.41, 20.84, and 17.25 for straight-take, fake-take, real-take (hires), and real-take (applicants), respectively. The means for the straight-takers/incumbents and real-take applicants were very similar. The mean was highest for those incumbents instructed to fake, indicating that maybe individuals won't fake (as much) unless told to do so. However, those applicants that wound up getting job offers did obtain higher mean scores than those not hired, or straight-takers.

Overall, there appears to be a need for further investigation into biodata validity and a comparison of predictive and concurrent validities on this measure.

Situational Judgment Tests

Weekley, J. and Jones, C. (1999). Further studies of situational tests. Personnel Psychology, 52, 679-699.

The authors reported results of two situational judgment tests used over a number of organizations. Results indicated that SJT scores were related to performance, cognitive ability, and experience (the weighted mean validities across the studies were .19, .45, and .20, respectively). However the weighted average correlation between SJT scores and tenure was .02, indicating (as mentioned by Quinones et al (1995)) that experience may be multifaceted. These results are important in that they suggest that SJT's may be a feasible alternative to cognitive ability tests. However, the authors caution that additional research needs to be conducted on SJT's, due to the possibility that they may not accurately predict applicant performance. In the words of the authors, "Research on SJTs using predictive validation designs is sorely needed."

McDaniel, M. and Nguyen, N. (2001). Situational judgment tests: A review of practice and constructs assessed. International Journal of Selection and Assessment, 9, 103-113.

The authors review the history of Situational Judgment Tests and make suggestions for future research directions. The authors cite previous research (McDaniel et al, in press), which found that SJT's predict job performance ($\rho = .34$) and are correlated with general cognitive ability ($r = .36$). The authors suggest that future research on SJT's is needed. Possible avenues for research include:

- Development of methodologies for targeting specific constructs
    - Better understanding of what constructs SJT's are measuring.
- Determine how item characteristics influence validity and adverse impact
- Determine the extent to which SJT's are faking resistant

McDaniel, M.; Morgeson, F.; Finnegan, E.; Campion, M.; Braverman. E. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. Journal of Applied Psychology, 86, 730-740.

This meta-analysis examined (among other things) the criterion related validity of SJT's and their relationship with cognitive ability and performance. The authors also tested possible moderators of the SJT-performance relationship, including the possible effect of using a predictive versus concurrent validity design. Results indicated that the mean validity of predictive designs were smaller than those with concurrent designs (.18, .35, respectively). However, the authors note that accurate conclusions cannot be drawn based on these results, because only six predictive studies (with participants) were included in the analysis (compared to 96 concurrent studies with 10,294). The authors state "Conclusions on these moderators should await the accumulation of more data and a replication with hierarchical extension of the current meta-analysis."

General Information Concerning Non-Cognitive Measures

Roth, P.; Bobko, P.; Switzer, F.; and Dean, M. (2001). Prior selection causes biased estimated of standardized ethnic group differences: simulation and analysis. Personnel Psychology, 54, 591-617.

The authors examined how prior selection on a first predictor (usually cognitive ability) affected the observed *ds* for second predictors (biodata, personality, situational judgment). Results of a Monte Carlo simulation revealed that the observed *ds* on the second predictor were underestimated. The "downward bias" in the observed standardized ethnic group differences ranged from 30-70%, in the presence of low selection ratios, high standardized ethnic group differences on the screening predictor, and when the first and second predictor correlated higher than .30. The authors suggest that further research should take into account range restriction when designing studies that look at ethnic group differences. This suggests research on unrestricted samples on such measures as SJTs, biodata, etc, is needed

Guion, R. (1990). Personnel assessment, selection, and placement. In M. Dunnette and L. Hough (eds.) Handbook of Industrial and Organizational Psychology, Palo Alto, Consulting Psychologists Press.

The author points out the much of the research on personality measures is flawed, with too many concurrent rather than predictive designs and too little replication. Guion also suggests that more research exploring the use of more narrow work related measures as opposed to the broad personality items we have today. Concerning biodata, he suggests the use of Hough's (1984) Accomplishment record as a biodata measure. However, Guion does not believe background items assessing years of experience is a valid predictor of future performance.

Schmitt, N.; Gooding, R.; Noe, R.; and Kirsch, M. (1984). Metaanalysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37, 407-422.

In a meta-analysis of validation studies the authors found that concurrent validation designs and predictive validation designs produce similar validity coefficients, however both of these designs produce coefficients higher than those produced in a predictive design that includes use of the selection instrument. The mean validity coefficients are .341, .296, and .259 for concurrent, predictive, and predictive with selection, respectively. In addition examination of type of predictor test indicated that personality tests produced the lowest mean validity coefficients, .149 while supervisor/peer evaluations and assessment centers produced the highest validity coefficients, .427 and .407 respectively. The mean validity coefficient for biodata measures was .243.

Barrett, G.; Phillips, J.; and Alexander, R. (1981). Concurrent and predictive validity designs: A critical reanalysis. Journal of Applied Psychology, 66, 1-6.

Barrett and colleagues address the criticism that concurrent designs are not accurate, and predictive designs are therefore more valid. The authors address the issues of "missing persons", restriction of range, and motivational and demographic differences of incumbents versus applicants. They conclude that predictive validity designs and concurrent validity designs are equivalent, and that criticisms (such as restriction of range) are present in both types of designs.