1997

The Validity of the B3R Entry-Level Examination for Fire Services as a Predictor for Future Job Performance

Terrance W. Gaylord '97
Illinois Wesleyan University

Running Head: Predictor for Future Job Performance

The Validity of the B3R Entry-Level Examination for Fire Services as a Predictor for Future Job Performance

Terrance W. Gaylord

Illinois Wesleyan University

Abstract

The use of aptitude tests by public and private organizations has drawn attention from the scientific community as well as the courtroom. While theorists have argued that aptitude tests are not valid, several research studies have indicated otherwise. The current study focused on the B3R Entry-Level Examination for Fire Services, an aptitude test administered by a growing number of municipalities nationwide for selection purposes. To test the predictive ability of the exam, a criterion-related validation approach was initiated. A total of eight subjects were administered the B3R exam which was accompanied with a 21-item job performance evaluation of the employees. A correlation between the two scores resulted in less than significant results.

The Validity of the B3R Entry-Level Examination for Fire Services as a Predictor for Future Job Performance

The use of aptitude tests by organizations to measure skills and predict future job performance in order to properly select employees has been subject to criticism for decades. Generally, aptitude measures a prospective employee's specific abilities related to reading comprehension, mathematical computation, and analytical reasoning. According to Adams (1989), "an aptitude developed for a given occupation will include the several skills and abilities that have been identified from a job analysis" (p. 16). Whether or not these aptitude tests provide meaningful, reliable, and predictive results remains a controversy within the scientific community. The present study attempts to determine the predictive validity of the entry-level examination for Fire Services developed by International Personnel Management Association (IPMA). The entry-level examination for Fire Services, referred to as the B3R, tests potential fireman on a number of intelligence skills ranging from situational judgment to vocabulary, most of which pertain to firefighting.

Munsterberg's work with the testing of railroad engineers in 1912 may be considered the originator of test usage for personnel job selection though much time elapsed between his work and the influx of studies conducted in the United States. During the first World War, psychologists were asked to join an effort to separate enlisted men into either officer or infantry positions. The military personnel were required to complete either the Army Alpha(literacy exam) or the Army Beta(illiteracy exam) which

would yield cognitive ability scores. Based on their performance on the tests, the recruits would be placed in their prospective positions (Hothersall, 1995). Following the first World War, a sudden increase of testing on part of organizations and companies began to emerge. However, this course of personnel selection by means of intelligence testing has faced extreme scrutiny on the grounds that the procedure did not provide valid predictive results. Thousands of validation studies have been conducted over the past fifty years in an attempt to discover the actual predictability within the realm of intelligence testing.

Ghiselli (1973) conducted hundreds of validation studies over the course of his scientific career. The focus of Ghiselli's studies were on general "cognitive ability" tests, more commonly called intelligence tests. Two characteristics which generally hold true with these types of exams are as follow: (1) Subjects are tested on several specific aptitudes such as verbal aptitudes, quantitative aptitudes, technical aptitudes, and sometimes even three-dimensional spatial problems and arithmetic reasoning (2) general cognitive ability tests may be used for a variety of jobs whether it be for a clerical position or a managerial position.

Ghiselli (1973) completed validation studies on twenty-one types of jobs including clerical positions, sales, trades and crafts, and several others. The correlations between the general cognitive ability tests and the measure of overall success(which varied between jobs but typically was based on supervisor ratings) provided significant results favoring the validation of

predictability with the general cognitive ability exams. Ghiselli discovered that as the validity of general cognitive ability predicting performance rating decreases, the complexity of the occupation decreases as well. In other words, general cognitive ability tests are apt to predict the performance of occupations which require a vast amount of different and complex skills than those occupations with minimal required skills. Despite these odds, even those occupations with the least amount of complexity in Ghiselli's study managed to produce significant positive correlations (1973). On top of Ghiselli's work, a criterion-related validation study supported the use of verbal aptitude tests in hospital settings.

A study by Distefano and Pryer (1985) focused on verbal tests administered by three hospitals to enhance their selection procedure of staff members. The criteria they tested the verbal scores against were rating scales on overall performance which were completed by staff supervisors. Distefano and Pryer obtained positive correlations between the two variables at all three hospitals thus concluding that the verbal test served as a valid predictor of future personnel performance.

While Distefano and Pryer (1985) found significant results, some researchers may argue that subjective evaluation measures do not provide reliable results as might more objective measures. Nathan and Alexander (1988) focused on five different criteria with which to correlate clerical scores on general cognitive ability scores. Scores on general cognitive ability exams were correlated with supervisor ratings, supervisor rankings, work

samples, production quantity, and production quality versus the scores obtained from the subjects. As a result, high validities were accumulated on all but one of the criteria-based measures. Significant positive correlations were found with the ratings, rankings, samples, and work quantity. Nathan and Alexander (1988) concluded that "no support for the assumption that 'objective' measures of performance are more predictable than 'subjective evaluation, at least not for clerical jobs" (p. 530). One of the possibilities why a positive correlation was not found between scores and production quality may have been that the overall quality of the finished products would more often than not be out of the control of the employee (1988).

Although several studies support the validity of general cognitive ability tests and their ability to predict future employee performance, many researchers remain skeptical about actually using these results to choose candidates for employment. McClelland (1973) argues that intelligence does not provide substantial support of the current hypothesis stating that aptitude tests are successful predictors of future job performance. In other words, he argues that psychologists are incapable of proving that intelligence is the reason why people are successful at work. Due to the fact that correlations do not imply causation, McClelland suggests criteria other than intelligence, such as motivation, may contribute to high performance. McClelland criticizes Thorndike and Hagen's 1959 study in which they obtained some 12,000 correlations between aptitude test scores and other future performance variable.

Thorndike and Hagen concluded "that the number of significant correlations did not exceed what would be expected by chance" (McClelland, 1973, p. 3). Furthermore, McClelland continues his critique on test validation with a focus on Ghiselli's study. McClelland claims that Ghiselli fails to cite his sources, nor does he accurately describe how he measured job proficiency in his correlations.

If general cognitive ability tests should not be used as predictors of future performance, as supported by McClelland, what should employers rely on? Lavigna (1992) approached this question with a study testing the validity of background characteristics and their potential to predict the overall performance of applicants. His study focused on six characteristics including undergraduate grade point average (GPA), degree level (undergraduate or graduate), academic quality of college/university attended, major field of study, professional work experience before being hired, and recruitment source (campus interview or direct application. Lavigna tested these characteristics with performance appraisal (PA) scores that employees received during the first couple of years with their respective organizations. As a result, grades (GPA) were the only variable that resulted in a positive correlation with the PA scores. Lavigna concluded that "the results of the background characteristics studied are not significant predictors of early career performance" (Lavigna, 1992, p. 356).

The overview of past research on the disputed success of general cognitive ability exams leads to the current study on an

aptitude test which tests specific skills. The only difference between general cognitive ability exams and specific aptitude exams is the titles the exams are given and the content of the questions. Both exams test individuals on skills such as reading comprehension, vocabulary, and mathematical computations. In a general cognitive ability exam, a reading comprehension task may include a paragraph about rock climbing followed with questions inquiring the nature of the passage. In a specific aptitude test, the same format in questioning is followed, however, the questions are relative to the specific job at hand. The results of extensive research on general cognitive ability exams is thus applicable to recent research on specific aptitude exams. The current study examined the predictive validity of the B3R, an aptitude test designed specifically with content associated with the duties of a firefighter.

The B3R Entry-Level Examination for Fire Service is not considered a general cognitive ability test. The B3R is a job specific aptitude "related to one's chance of learning to perform optimally in the target job" (Adams, 1989, p. 17). Although this exam tests its applicants on skills exactly as do general cognitive ability exams such as reading comprehension and numerical ability, it is unique due to its focus on the skills and abilities which have been identified by a job analysis. Job analyses are conducted by a number of experts within a given field who evaluate all of the questions to be included in an aptitude test. The experts must agree that the selected items for the exam are related to the job the aptitude test is

specifically designed for. Specific ability tests have not been deemed valid predictors of future performance as have general ability tests. According to Arvey (1986), there is primarily one reason for this. General cognitive ability tests focus on those of an individual's abilities that manage all other specific abilities. In other words, a person who has high reading comprehension abilities will be able to apply this ability to a several specific literary items, i.e. newspaper, autobiography, or work-related reports. Whether or not specific aptitude exams maintain the power to predict was examined by Rafilson & Sison (1996).

Rafilson and Sison (1996) analyzed seven criterion-related validation studies associated with police departments nationwide. The results obtained from these studies were exceptionally supportive of the predictive validity of the National Police Officer Selection Test (POST) when those scores were correlated with police academy test scores and with job performance evaluation scores. The POST is a police officer entry-level examination administered by a number of police departments for selection purposes. One of the studies examined by Rafilson and Sison viewed the relationship between officers' scores on the POST and the scores they received on academy tests towards the end of their training program. The positive correlation between the two scores was significant and supports the claim that the POST is an indicator of police academy performance. Another study analyzed by Rafilson and Sison focused on the subjective criterion-related validity of the POST. A police department in

Texas administered the POST to 246 law enforcement incumbents prior to evaluating the job performance of the police officers. Again, a significant positive correlation existed between the two variables significantly. "These results further support prediction from POST scores of the job performance of law enforcement officers in a variety of settings" (Rafilson & Sison, 1996, p. 174).

More support on the criterion-related validity of aptitude tests was found in a study by Cesare, Blankenship, Giannetto, and Mandel (1993). They focused on the entry-level examination for Eligibility Technicians which tests applicants on reading comprehension skills, interpersonal skills, and more. "The singular intent of the test is to distinguish between those candidates who demonstrate a satisfactory knowledge of the minimum qualifications necessary for suitable performance as an Eligibility Technician" (Cesare et al, 1993, p. 112). Both subjective and objective criterion-related measures were correlated with the 3,268 results on the entry-level exam. The subjective criterion measure was a 22-item performance appraisal rating employees on a scale of one to five. The objective criterion measure was the tenure of each employee, or the length of employment. As a result of the study, "the written test was shown to be a successful predictor of subsequent job performance; individuals who performed well on the test, in turn performed better on the job then those who scored poorly on the test" (Cesare et al, 1993, p. 120).

The present study focuses on the B3R Entry-Level Examination for Fire Services. This aptitude test is administered by municipalities nationwide as a pre-screening tool for hiring processes. The study attempts to support the hypothesis that the B3R exam is a valid predictor of future job performance through the means of a criterion-related validation study.

Method

Participants

The participants for this study are eight current employees of the Fire Department serving the Town of Normal. The number of participants have been kept to a minimum for one primary reason; the length of their employment. When conducting a criterion-related validation study for concurrent validity, it is essential to test incumbents who have no more than six months to one year of job experience (Adams, 1989, p. 19). Of the eight subjects, only two of them exceed this limit by a maximum of only two months. All of the subjects are white males with no more than one year and two months of experience with the department.

Materials

Participants were given the B3R Entry-Level Examination for Fire Services. This exam was developed by International Personnel Management Association (IPMA) in 1995. The exam contains ninety items which test applicants on quantitative, analytical, and verbal skills associated with firefighting. IPMA rents tests and testing materials to municipalities interested in using the exam for a period of sixty days. Evaluation forms to

be completed by the participants' supervisors were originally created by the Town of Normal. Questions, rating scales, and format of the evaluation have been revised to better suit the study.

Procedure

In order to determine whether or not this had an impact on the participants' knowledge and test scores, a correlation between scores and length of employment was calculated. The eight subjects were given the B3R Entry-level Examination for Fire Services prepared by International Personnel Management Association (IPMA). Over a period of four days, the eight participants completed the exam in one of the testing rooms located in the basement of the Normal Fire Department. One week after the exams were completed, the direct supervisor of the participants was given performance evaluation forms which were completed within a week. A Pearson Product-Moment Coefficient will be calculated to determine whether or not a positive, significant relationship does exist between the two scores.

Results

The scores on the B3R ranged from a 58 to an 87. The highest achievable score on the B3R is a 90. The mean score of the participants was a 74.8. The mean on the evaluations was a 108.6; the highest achievable score on the evaluation was a 147 (21 questions multiplied by 7, the possible high for each question).

Table 1

Participant Scores on the Two Measured Variables

Participant B3R Scores	Participant Evaluation Scores
87	103
80	131
78	139
77	86
76	107
73	94
70	107
58	102

A Pearson Product-Moment Coefficient was calculated using the eight sets or pairs of scores received from the eight participants. The first score in each set represents the score the subject earned on the written B3R exam. The second score represents the total score earned on the twenty-one question evaluation form completed by the direct supervisor. A Pearson r score of r=.22 was obtained. A two-tailed t-test for significance resulted in non-significant results at the .05 level. Based on these results, minimal predictions if any could be made between the two scores. In other words, high scores on the B3R exam do not necessarily associate with high scores on the performance evaluation. However, do to the small sample size, this matter needs to be evaluated more closely.

Discussion

Despite a failure to produce significant findings in this study, a careful observation and analysis of the scores indicates a definite trend in the results. Four pairs of scores, (73-94),

(76, 107), (80, 131), and (78, 139) represent a positive linear relationship, as indicated in Figure 1. The participants who achieved scores of (70, 107), and (77, 86) are also extremely close to this trend only with a lower exam score and a lower evaluation score respectively. The outlyers of the sample pool were the participants who scored a (58, 102) and a (87, 103) for the study. The two outlyers in the results have been marked with an X on the graph. A closer look reveals the participants seemingly low evaluation score in comparison to the score of an 87, resulted due to two below average scores for attendance on the evaluation. This category is not targeted in anyway in the B3R exam. In other words, there is no way to predict the participants ability to show up to work consistently based on a test score. This result should have no bearing on the overall results of the study. Interestingly, the participant who scored a mere 58 on the exam, missed a number of questions on both safety and public relations, 30%. The evaluation of this participant shows that the lowest scores earned were in those same categories, safety and public relations. This comparison supports the predictability of the exam. Overall, it appears that either an addition of participants or an elimination of the two extreme scores would result in significant findings.

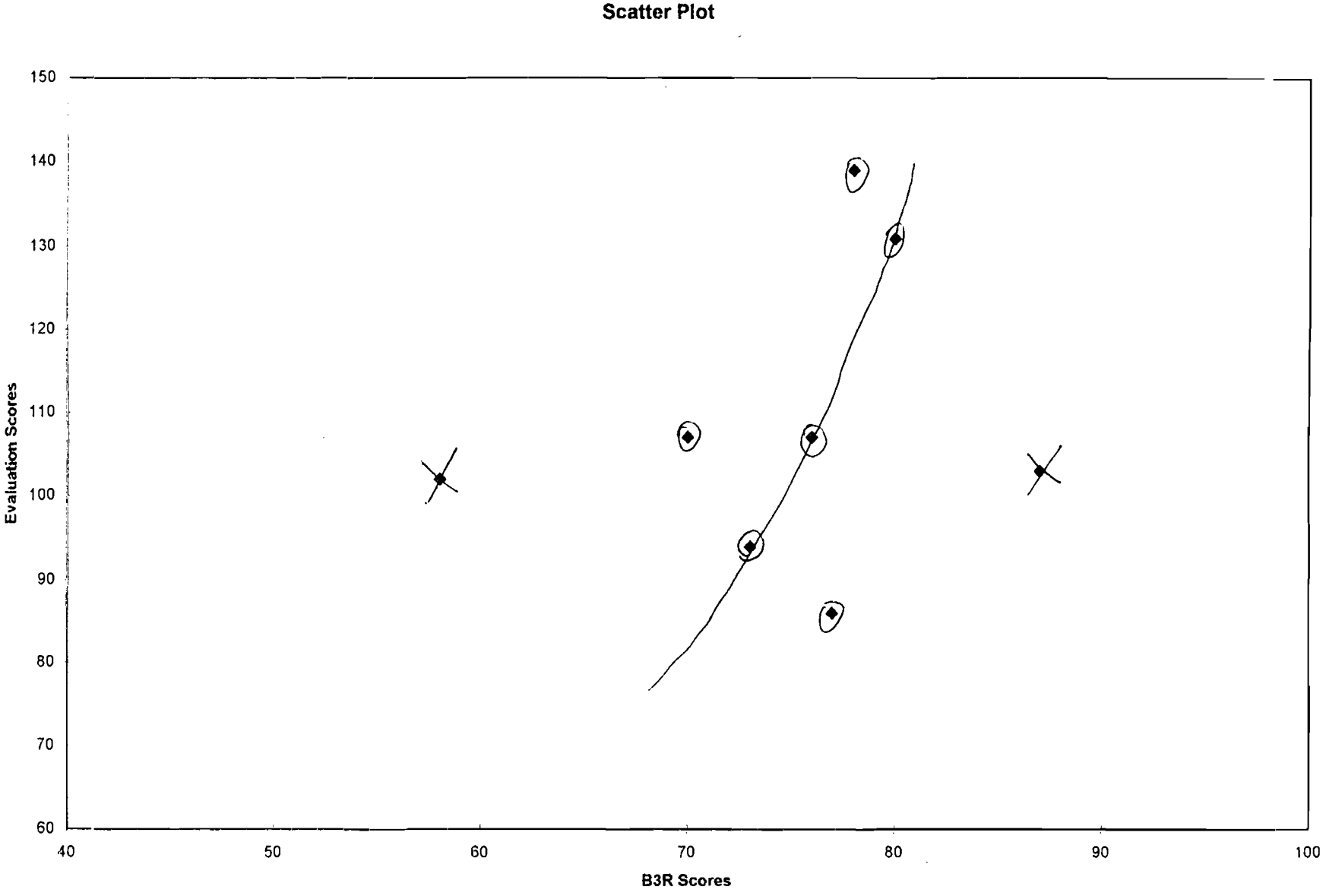
There are several problems with the current study which may have contributed to the lack of significant findings. First of all, due to the low participant pool, the two outlyers managed to disrupt the overall results. The low score of a 58 could have resulted from several factors. The participants had no incentive

for taking the exam. They were not paid to take the exam and they were not offered any promotional rewards for their efforts. On top of this, their scores were held confidential which may have decreased any motivation for taking the exam. One last problem may have been the fact that the participants were undergoing other paper and pencil testing for the Normal Fire Department at the same tie they participated in the study. They may have been mentally fatigued as a result and thus scored lower on the exam than their potential, i.e. a 58. One must also consider that these scores are indicative of their actual intelligence and ability on tests whereas a larger participant pool is needed to overcome extreme scores.

A replication of this study should include some of the following improvements. For one, the participant pool needs to be increased for obvious reasons stated previously. Next, the participants, in order to increase their motivation and effort, should be offered a reward of some sort for outstanding performance i.e. $15 for a score over 80. Finally, because raters judge behavior differently, each participant should be scored on their evaluations by more than on supervisor. Then the two scores should be averaged out which was not what happened in this study.

Figure Caption

Figure 1. Scatter plot of the two variables including B3R scores along the x-axis and evaluation scores along the y-axis.


Scatter Plot
150
140
130
120
110
100
90
80
70
60
Evaluation Scores
40
50
60
70
80
90
100
B3R Scores

Adams, Joanne. (1989). Handbook on test administration. Alexandria, Virginia: IPMA.

Arvey, Richard D. (1986). General Ability in employment: a discussion. Journal of Vocational Behavior, 29, 415-420.

Barret, Gerald V., & Depinet, Robert L. (1991). A reconsideration of testing competence rather than for intelligence. American Psychologist, 46, 1012-1024.

Cesare, Steven J., Blankenship, Mark H., Giannetto, Patrick W., & Mandel, Mark Z. (1993). A predictive validation study of the methods used to select eligibility technicians.

Distefano, M. K., & Pryer, Margaret W. (1985). Verbal selection test and work performance validity with aides from three psychiatric hospitals. Psychological Reports, 56, 811-815.

Ghiselli, Edwin E. (1973). The validity of aptitude tests in personnel selection. Personnel Psychology, 26, 461-477.

Hothersall, David (1995). History of Psychology. New York: McGraw-Hill.

Hunter, John E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. Journal of Vocational Behavior, 29, 340-362.

Lavigna, Robert J. (1992). Predicting job performance from background characteristics: more evidence from the public sector. Public Personnel Management, 21, 347-361.

McClelland, D. C. (1973). Testing for competence rather than for "intelligence." American Psychologist, 41, 1-14.

Nathan, B. R., & Alexander, R. A. (1988). A comparison of criteria for test validation: a meta-anaylitic investigation. Personnel Psychology, 41, 517-535.

Rafilson, Fred, & Sison, Ray. (1996). Seven criterion-related validity studies conducted with the national police officer selection test. Psychological Reports, 78, 163-176.