


# AllBusiness

 Print Page

## Fire service testing in a litigious environment: a case history.

Since the mid-1970s, the city of St. Louis has encountered frequent litigation and court involvement surrounding the testing procedures used to evaluate candidates for promotion in the fire service. During this period, the city has expended a considerable amount of time, effort, and money to develop valid and defensible methods of testing. In light of this experience, we have decided to review the past history to see if any conclusions can be drawn that would be helpful to other organizations. This article contains a brief history of the city's fire service testing and litigation, and a description of the current testing model used by the city. It also addresses the use of various testing components and their impact on all candidates. Finally, some practical considerations are discussed.

In 1974, the percentage of African-American employees in the St. Louis Fire Department was quite low. Of approximately 1,000 uniformed personnel in the department, only 110 (11%) were black. Of the 180 persons in the rank of Fire Captain, only 4 (2%) were black. No black employee had held a position above the rank of Fire Captain in the history of the department. This was in a city with a more than 40% black population according to 1970 census figures. It was obvious that this staffing pattern was unacceptable. It was not, however, an easy problem to solve.

Griggs v. Duke Power<sup>(1)</sup> was only a few years old in 1974, and fair employment case law had still not clearly defined terms such as "adverse impact." Additionally, the federal court system was continually evolving its own standards for what constituted acceptable test validation. Criterion-related validity was still regarded as the norm by the psychological testing community, but problems such as bias in performance ratings and claims of "differential validity" forced the city of St. Louis, as well as many other jurisdictions, to consider other validation methods.

The use of content validity as a means of validating selection instruments was just beginning to be accepted by the testing community. Voluntary affirmative action programs were virtually unheard of at that time, and the strong civil service rules and regulations in effect in St. Louis made it difficult to consider relaxing requirements for strict rank-ordering of candidates on eligible lists based on test scores. There was also a strong reliance on traditional civil service testing procedures such as multiple-choice written tests. As recognition of the problems inherent in paper and pencil testing increased, it became apparent that changes in the city's civil service testing procedures were necessary. As a result, innovations such as fire scene simulations and the use of assessment center exercises evolved.

A dramatic increase in minority employment has occurred in the department since 1974. In 1994, approximately 280 (40%) of the 700 uniformed personnel in the department were black. In addition, 39 (32%) of 123 Fire Captains, 10 (50%) of 20 Battalion Fire Chiefs, and 3 (75%) of 4 Deputy Fire Chiefs were black. The most recent Census Bureau data (1990) reveals that the population of the city of St. Louis is 50.9% caucasian and 47.5% African American. Other minority groups, although growing presently, were not as widely representative in the community during the period in question.

There are several reasons for the dramatic shift in the fire department's minority representation. In part the increase was due to a consent decree for the entry-level firefighter position, in effect since the mid-1970s. The decree calls for hiring candidates on a 50% white and 50% black basis. The increase has also been due to a consent decree for the Fire Captain position, in effect from 1980 to 1983. This decree called for vacancies to be filled on the basis of one black candidate for each two white candidates being promoted. One additional reason for the increase, especially at the levels of Fire Captain and Battalion Fire Chief, is the dramatic change in civil service testing procedures in St. Louis since 1983.

Testing underwent many changes since the early 1970s. Today, many public jurisdictions use content validity and innovative testing procedures such as job simulations. Perspectives on the changes differ among the groups involved in the long series of events. To minority firefighters the progress has seemed agonizingly slow. On the other hand, this was a frustrating period for white firefighters. Many felt they were victims of "reverse discrimination," that racial balance was the only criterion of acceptability, and "validity" was only a ruse for social engineering. However, in retrospect, as frustrating as this period was for those involved, the accomplishments seem remarkable. The changes instituted preserved the excellent reputation of the St. Louis Fire Department, while bringing minority staffing levels to a point that approximated population figures of the city as a whole.

## Early Litigation

### (1974 and 1979 Fire Captain Examinations)

The purpose of this section is to review the early litigation against the city of St. Louis. This litigation resulted in changes to the city's traditional testing procedures and ultimately led to the general testing model that is now used for four separate promotional examinations in the fire department.

Litigation began with a 1974 testing procedure for Fire Captain consisting of a multiple-choice written test measuring technical knowledge, an experience and training evaluation measuring length of service, and a service rating evaluating firefighter job performance. The written test and experience and training evaluation each had a weight of 45%. The service rating had a weight of 10%. The examination process also included a working test period (probationary period) as a pass/fail component to evaluate supervisory skills. This testing procedure had an adverse impact against black candidates, and its validity was challenged by a group of black firefighters (Firefighters Institute for Racial Equality (F.I.R.E.)), and by the U.S. Justice Department.(2)

The Federal District Court ruled in favor of the city, but the ruling was overturned by the Eighth Circuit Court of Appeals.(3) The Appeals Court said that the captain's exam had admittedly failed to test a major aspect of the job that separated a Fire Captain from a Firefighter, that being supervisory positions. The Appeals Court said that a working test period was not an acceptable way to evaluate supervisory skills because candidates not selected for the working test period obviously had no way of demonstrating their skills. In its decision, the Court referred to the "assessment center" as an excellent way to evaluate candidates for supervisory positions. The city had considered using an assessment center but decided against it because of the large number of candidates to be tested and the substantial costs. In December 1978 the Appeals Court ordered the city to come forth with a valid examination by January 1, 1979. (4) (Fortunately, the city had already begun working on a new examination and was able to meet this deadline).

The 1979 Fire Captain examination consisted of a written test measuring technical fire fighting knowledge weighted at 30% and an assessment center measuring both technical firefighting knowledge and supervisory skills weighted at 70%. The assessment center consisted of three exercises: a "fire scene simulation" where candidates viewed slides of a fire and wrote their observations and the orders they would give if they were a Fire Captain in charge of fighting this fire, a "training simulation" where each candidate was given informational material and was required to prepare and present a lecture similar to one that might be given at the fire house, and an "interview simulation" where each candidate played the role of a Fire Captain and had to deal with a person playing the role of a firefighter who was having a personal problem with another firefighter.

White candidates scored significantly higher than blacks on the written test (White mean = 98.04, Black mean = 89.54,  $p$  [less than] .05). However, there was no statistically significant difference between white and black candidates on the assessment center. When scores from the two components were combined according to the published weights, there was an adverse impact against the black candidates.

The validity of the exam was contested by F.I.R.E. and the U.S. Justice Department.(5) The Federal District Court ruled in favor of the city but was reversed by the Eighth Circuit Court of Appeals.(6) The Appeals Court said that the written (multiple-choice) test did not meet the requirements for content validity under the Uniform Guidelines (1978).(7) The Court ruled that since the written test was a weighted component used to assist in ranking candidates, "empirical evidence" was required. In discussing the assessment center, the Appeals Court mainly criticized the fire scene simulation. The Appeals Court referred to this as a "paper-and-pencil test which is far removed from the content and context of the candidate's actual work behavior," and again said that "empirical evidence" was necessary to document its validity.

In 1980, the Appeals Court ordered the city to rank-order the candidates based on the assessment center scores. Furthermore, vacancies were to be filled on the basis of one black for each two white firefighters being promoted. The Court also ordered that all the parties involved in the litigation work together to develop a testing procedure in compliance with the Uniform Guidelines. The Court also required the city to pay for experts to assist each of the groups in preparing the examination. The testing procedure (model) developed by this group of experts will be the primary focus of the remainder of this article.

## Testing Procedure

As a result of the 1980 Federal Appeals Court order, the city needed to develop a new and valid testing process for Fire Captain. The city was now obligated to work with a committee of experts to prepare a mutually acceptable, valid examination. This Test Development Committee consisted of experts representing Local 73 of the International Association of Fire Fighters (I.A.F.F.), the Firefighters Institute for Racial Equality (F.I.R.E.), the U.S. Justice Department, and a representative from the city's personnel department. The testing procedure developed by this

committee was content-validated and consisted of a multiple-choice written test measuring basic technical knowledge, a fire scene simulation measuring advanced technical knowledge, and an assessment center measuring supervisory and administrative skills. The written test and fire scene simulation were used on a pass/fail basis. The assessment center had a weight of 100%. (This weighting was influenced by the earlier court action, which ordered a similar weighting when the assessment center included the fire scene simulation). This general testing model, with some minor modifications, has now been used on four separate occasions to evaluate candidates for promotion to Fire Captain (3 occasions) and Battalion Fire Chief (1 occasion). The exams covered in this article include: 1983 Fire Captain (83 FC), 1990 Fire Captain (90 FC), 1994 Fire Captain (94 FC), and 1986 Battalion Fire Chief (86 BFC).

#### Written Test

Two different multiple-choice written tests were used during the researched period. One was developed by the Test Development Committee (83 FC and 90 FC). The other was purchased (rented) from a test publishing organization (86 BFC). The written tests measured knowledge of basic firefighting procedures and techniques; knowledge of St. Louis Fire Department rules, regulations, and procedures; and, in some cases, knowledge of supervisory and management principles and practices. In all cases, the written tests were used on a pass/fail basis and were not weighted components. Candidates had to receive a passing score to proceed to the next exam component. The 94 FC exam did not have a written test component.

#### Fire Scene

The fire scene component was a paper-and-pencil simulation in which fire scenarios were described in a written text and shown in drawings and/or photos. Candidates were instructed to write descriptions of the actions they would take as the officer in charge of each situation, including orders they would give, personnel assignments they would make, and equipment and tools they would use. Candidates also used a "street map diagram" to show where they would place their apparatus, the size and location of hose lines and ladders, and the point where they would attack the fire. In all but the 94 FC exam, the fire scene was used on a pass/fail basis and was not a weighted component. On the 94 FC exam, the fire scene had a weight of 10%. Three different consultants were used to develop the fire scene simulations.

#### Assessment Center

The assessment center consisted of job simulations designed to evaluate supervisory, managerial, and administrative skills. The simulations included several different types of exercises such as in-baskets, coaching and counseling sessions, interview simulations, leaderless group discussions, training simulations, briefings, oral presentations, and accomplishments surveys. Depending upon the promotional position, the number of exercises ranged from two to four. Two different consultants were used to develop assessment centers (one was used twice). One assessment center (86 BFC) was developed in-house. For all the examinations, the assessment center was used as the sole or primary basis for ranking candidates on the eligibility list.

#### Uses and Results of Exam Components

Tables 1-3 are included to illustrate how each exam component was used and to provide component results by race. Table 1 shows the number and percentage of candidates eliminated on each component. (The assessment center is not included because only one candidate has been eliminated over the four exams.) Table 2 provides a comparison of success rates using the "four-fifths (80%) rule." (The assessment center is not included because all candidates taking the assessment center are typically placed on the eligible list and are ultimately promoted unless they leave the fire service while on the list.) Table 3 compares mean scores by race on the exam components.

#### Written Test

For the three exams that included a written test component, the written test was used to screen out approximately 18% of the candidates (please refer to Table 1). The rationale for the low screen-out rate was that the written test could only be adequately used to determine whether candidates possessed the basic knowledge required for the job. When comparing the passing rates by race using the "four-fifths (80%) rule," none of the written tests had an adverse impact (please see Table 2). When comparing mean differences by race, whites scored significantly higher than blacks on all of the written tests (please see Table 3).

#### Fire Scene

The fire scene was given a more important role in screening candidates than the written test because the Test Development Committee felt that the fire scene more closely replicated the job and, consequently, was a better measure of firefighting skill. For the four exams, the fire scene was used to screen out approximately 69% of the

candidates (Table 1). When comparing the passing rates by race using the "four-fifths (80%) rule," the fire scene had an adverse impact against white candidates on three exams (83FC, 90FC, and 86BFC)(Table 2). On two of these occasions the impact was substantial (83FC = 48%, 86BFC = 46%). When comparing mean differences by race, blacks scored higher than whites on all fire scenes. However, on only one of the fire scenes was the difference statistically significant (86BFC) (Table 3).

#### Assessment Center

For the four exams, the assessment center was used as the sole (or primary) basis for ranking candidates on an eligible list and was not typically used to eliminate candidates from consideration. It was also typical for all candidates on an eligible list to be promoted prior to the expiration of the list. The decision to use the assessment center as the sole basis to rank candidates was based on the rationale that the written test and fire scene had successfully eliminated most of the candidates whose job knowledge and abilities were deficient. Among those still in the pool, the differences in performance among candidates would depend on the behaviors measured in the assessment center. Since candidates were not screened out on the assessment center and since all available candidates on an eligible list were ultimately promoted, a comparison of success rates by race is not relevant. When comparing mean differences by race, the results have been mixed. On two of the assessment centers (83 FC and 94 FC), whites scored higher than blacks, but the differences were not statistically significant. On the other two assessment centers (90 FC and 86 BFC), blacks scored significantly higher than whites (Table 3).

#### Discussion and Conclusions

The testing model described in this article has achieved considerable success at minimizing adverse impact against black candidates in the 15 years since its implementation in St. Louis. However, there are some areas of concern that should be reviewed before the model is implemented elsewhere.

It has been our experience that whites have consistently scored higher than blacks on the written test components. Blacks have consistently scored higher than whites on the fire scene components. Assessment center results have been mixed with whites scoring higher than blacks on two occasions, and blacks scoring higher than whites on the other two occasions. Overall, however, the testing model has not produced a bottom-line adverse impact against black candidates on any of the four examinations discussed in this article.

Several factors have contributed to the overall absence of adverse impact against black candidates. First of all, we have set relatively low cut-off scores on the written test components. In addition, the written tests were used only on a pass or fail basis and were not used as weighted components to assist in ranking candidates on the eligible lists. Secondly, we have set relatively high cut-off scores on the fire scene simulations. The fire scenes were used as the primary screening devices to get into the assessment centers and, as with the written tests, were typically not used as weighted components. Finally, we have typically not screened out candidates on the assessment centers, but have used them as the primary basis for ranking candidates on the eligible lists. We have also traditionally promoted all available candidates on the lists.

Table 1

#### Use of Examination Components to Eliminate Candidates

##### WRITTEN TEST

##### Number and % of Candidates Eliminated

83 FC	83/320 = 26%
90 FC	31/287 = 11%
94 FC	NO WRITTEN TEST WAS USED
86 BFC	15/96 = 16%
TOTAL	129/703 = 18%

##### FIRE SCENE

##### Number and % of Candidates Eliminated

83 FC	149/197 = 76%
90 FC	146/233 = 63%
94 FC	177/251 = 71%
86 BFC	51/81 = 63%

TOTAL 523/762 = 69%

Though this has historically been our experience, it is difficult to project these same results into the future with certitude. Both blacks and whites are becoming more sophisticated in their methods of test preparation, and the consultants who prepare the various test components constantly change as a result of the city's competitive bidding process. In addition, because cut-off points on both the written and fire scene components are set to minimize adverse impact, greater weight is placed on the assessment center component as the primary ranking device for the examination. We have grown increasingly uncomfortable with this situation.

The use of the assessment center component as the primary ranking device in Fire Department promotional testing in St. Louis is historical in derivation. It stems from the almost universal adverse impact of paper and pencil tests against minorities, and from early criticisms against a locally developed fire scene simulation as being too similar to a paper-and-pencil test. Even when the court-appointed expert panel developed more acceptable written and fire scene components, the earlier approved practice of using the assessment center as the primary ranking device was continued. We believe that this practice needs to be reviewed, and that weights for the various examination components be based on a thorough job analysis.

Table 2

Comparison of Passing Rates by Race Using the "Four-Fifths (80% Rule)"

WRITTEN TEST

	Blacks Passing	Whites Passing	Impact (1)
83 FC	67% (41/61)	76% (196/259)	89%
90 FC	78% (63/81)	94% (193/206)	83%
94 FC	NO WRITTEN TEST WAS USED		
86 BFC	71% (17/24)	89% (64/72)	80%

1 In all cases, the impact was against black candidates

FIRE SCENE

	Blacks Passing	Whites Passing	Impact (1)
83 FC	43% (15/35)	20% (33/162)	48%
90 FC	45% (27/60)	35% (60/173)	78%
94 FC	32% (30/95)	28% (44/156)	89%
86 BFC	65% (11/17)	30% (19/64)	46%

2 In all cases, the impact was against white candidates

We believe that the use of a written test to measure job knowledge also needs to be reviewed. There are several reasons for this. First, paper-and-pencil tests historically have an adverse impact against blacks. Second, there is at least some duplication in terms of what is measured on the written test and fire scene simulation. Third, the number of candidates traditionally eliminated on the written test is so small that it would be more cost effective to let those candidates compete on the fire scene simulation.

A word of caution, however, before arbitrarily eliminating the written test component. The positions of Fire Captain and Battalion Fire Chief require technical knowledge in a number of areas. While a fire scene simulation may measure some of this knowledge, the focus may be on a specific and limited area. A written test, however, may more comprehensively assess technical job knowledge. Therefore, the decision on whether to eliminate the written test should be based on a job analysis as well as what factors are being measured by the other testing components.

[TABULAR DATA FOR TABLE 3 OMITTED]

One additional modification that may be worthy of review is the use of multiple hurdles. The City of St. Louis has traditionally used a multiple hurdles approach in which candidates had to achieve a passing score on one exam component to proceed to the next component. The advantage is that the number of candidates is reduced for the exam

components, which are most costly, time-consuming, and labor intensive to administer (e.g., assessment center). An alternative to this would be an approach where all candidates would complete the entire examination process. The advantage of this would be the elimination of the hurdles that are likely to result in adverse impact. This approach may also result in a higher level of acceptance among candidates since they would all be allowed to compete in the entire examination process.

In conclusion, while there are obvious legal and public relations problems associated with testing procedures that result in a disparate impact, the search for valid testing devices without adverse impact against minority groups may appear futile. The St. Louis experience suggests otherwise. While it is hazardous to generalize on the basis of one jurisdiction's experience, it is interesting to note that over four test administrations using different tests and testing consultants, the results were remarkably consistent. The tentative conclusion is that the testing procedures can make a difference in relative impact. What is promising is that a search for valid testing devices without adverse impact may be fruitful after all, with the consequent avoidance of court orders, bad publicity, and aggravated racial tensions. And, that seems to be a worthy goal for all human resources professionals.

The authors would like to thank Ed Knowles, Deane Looney, Beverly Ritter, Chris Bruening, Terry Dabrowski, Laura Stephens, and Joan Pynes for their assistance in making this paper possible.

#### Notes

1 Griggs v. Duke Power, 3 FEP 175 (1971)

2 F.I.R.E. v. City of St. Louis, 14 FEP 1473 (1976)

3 F.I.R.E. v. City of St. Louis, 14 FEP 1486 (1977)

4 F.I.R.E. v. City of St. Louis, 18 FEP 1083 (1978)

5 F.I.R.E. v. City of St. Louis, 19 FEP 1643 (1979)

6 F.I.R.E. v. City of St. Louis, 21 FEP 1140 (1980)

7 U.S. Equal Employment Opportunity Commission, U.S. Civil Service Commission, U.S. Department of Labor, and U.S. Department of Justice. 1978. Uniform Guidelines on Employee Selection Procedures. Federal Register, 1978, 43 (166) 38295-38309.

#### Authors

Gary M. Gebhart, Assistant Examination Manager City of St. Louis Department of Personnel Room 100, City Hall St. Louis, MO 63103

Gary M. Gebhart is the Assistant Recruitment and Examination Manager for the St. Louis City Personnel Department. He has over 15 years of experience in testing for fire service positions, including considerable experience with fire scene simulations and assessment centers. Mr. Gebhart holds an M.A. degree in Industrial-Organizational Psychology from Southern Illinois University at Edwardsville.

William C. Duffe, Director Of Personnel City of St. Louis Department of Personnel 1300 Convention Plaza St. Louis, MO 63103

William Duffe has served as Director of Personnel for the City of St. Louis since 1978 and has been active in IPMA at the Local and Regional Levels. He has an M.B.A. degree from St. Louis University and M.A. degree in economics from the University of Missouri St. Louis.

Roger A. McCurley, Recruitment & Examination Manager City of St. Louis Department of Personnel Room 100, City Hall St. Louis, MO 63103

Roger McCurley serves as Recruitment and Examination Manager for the City of St. Louis. He holds a Master of Public Administration degree from Southern Illinois University Edwardsville.