UNITED STATES DISTRICT COURT
DISTRICT OF MASSACHUSETTS

CIVIL ACTION NO. 07-11693-GAO

PEDRO LOPEZ, et al.,
Plaintiffs,

v.

CITY OF LAWRENCE, et al.,
Defendants.

FINDINGS OF FACTS, CONCLUSIONS OF LAW, AND ORDER FOR JUDGMENT
September 5, 2014

O'TOOLE, D.J.

I. Introduction

The individual plaintiffs are current or former police officers employed by the cities of Boston, Lawrence, Lowell, Methuen, Springfield, or Worcester, or by the Massachusetts Bay Transportation Authority ("MBTA"). Each of the plaintiffs is self-described as either African-American or Hispanic. Each took at least one civil service examination administered by the Human Resources Division ("HRD"), an agency of the Commonwealth of Massachusetts,[1] for promotion to the rank of sergeant within his or her respective police forces during the years 2005, 2006, 2007 or 2008, and, based largely on the resulting test scores, was not promoted or was promoted after white police officers from the same jurisdiction who took the same exam. By this lawsuit, the plaintiffs claim that the defendants' reliance on the HRD civil service exam in selecting candidates to promote resulted in unlawful "disparate impact" discrimination on the basis of race or ethnicity in violation of Title VII of the Civil Rights Act of 1964 and of Chapter

[1] HRD is an agency within the Executive Office of Administration and Finance.

151B of the General Laws of Massachusetts.[2] They seek injunctive and declaratory relief, as well as appropriate compensatory damages.

II. Proceedings

By a prior order , this Court bifurcated this action into two stages, the first to address the defendants' respective liability claims and a second, if necessary, to determine appropriate remedies. The plaintiffs' motion for class certification was denied as to the liability stage and denied without prejudice as to the potential remedial stage.

As originally brought, this action included claims against the Commonwealth of Massachusetts and HRD, which is charged under state law with the responsibility to prepare and administer written examinations for hiring and promotion to public employer positions subject to the State's civil service law. This Court had earlier denied the state defendants' motion to dismiss the claims against them, but that ruling was reversed on appeal. See Lopez v. Massachusetts, 588 F.3d 69, 90 (1st Cir. 2009). Consistent with the direction of the Court of Appeals, the state defendants were dismissed from the suit.

The liability phase of the case was tried against the municipal and MBTA defendants in a lengthy bench trial.[3] This memorandum and order resolves the factual and legal issues presented at trial.

[2] The parties agree that the legal analysis of a disparate impact claim of employment discrimination under Mass. Gen. Laws ch. 151B, § 4 is the same as analysis of such a claim under Title VII. See White v. Univ. of Massachusetts, 574 N.E.2d 356, 358 (Mass. 1991) ("The analysis of a discrimination claim is essentially the same under the State and Federal statutes."); see also Sullivan v. Liberty Mut. Ins. Co., 825 N.E.2d 522, 529 n.10 (Mass. 2005) (referring to federal standards for disparate impact claims under Massachusetts law). Accordingly, the discussion herein will focus on the well-developed principles of federal law, applicable to both the federal and state claims.

III. Legal Context

Before addressing and resolving disputed factual issues on the basis of the trial evidence, it will be useful to summarize pertinent principles of law.

A. Disparate Impact Discrimination Generally

Title VII of the Civil Rights Act of 1964 ("Title VII"), as amended, prohibits any employment practice that has "a disparate impact on the basis of race, color, religion, sex, or national origin." 42 U.S.C. § 2000e-2(k)(1)(A)(i). Unlike claims of disparate treatment, disparate impact claims do not require proof of an intent to discriminate. Ricci v. DeStefano, 557 U.S. 557, 577 (2009); Bradley v. City of Lynn, 443 F. Supp. 2d 145, 155 (D. Mass. 2006); see also, Sch. Comm. of Braintree v. Massachusetts Comm'n Against Discrimination, 386 N.E.2d 1251, 1254 (Mass. 1979) (addressing claims under Mass. Gen. Laws ch. 151B). The purpose of a disparate impact claim is to "'root[] out 'employment policies that are facially neutral in the treatment of different groups but that in fact fall more harshly on one group than another and cannot be justified by business necessity.'" Bradley, 443 F. Supp. 2d at 155 (quoting EEOC v. Steamship Clerks Union, Local 1066, 48 F.3d 594, 600-01 (1st Cir. 1995)).

There are potentially three steps involved in the proof of a disparate impact employment discrimination claim. First, a plaintiff has the initial burden of proving that a challenged employment practice has a disparate adverse impact on employees of a particular racial, ethnic or other protected group. Bradley, 443 F. Supp. 2d at 156 (citing 42 U.S.C. § 2000e-2(k)(1)(A)). Second, after a plaintiff has demonstrated such a disparate impact, the burden shifts to the employer to prove that the challenged practice is nonetheless "job-related and consistent with

[3] In addition to the municipal defendants and the MBTA itself, the plaintiffs have sued various officials of those entities. There is no need separately to address the individual defendants; their fates fall or rise with those of their respective entities.

business necessity." Id. at 157 (quoting Steamship Clerks, 48 F.3d at 601-02); see also 42 U.S.C. § 2000e-2(k)(1)(A)(i) (requiring employer "to demonstrate that the challenged practice is job related for the position in question and consistent with business necessity"); id. § 2000e(m) ("The term 'demonstrates' means meets the burdens of production and persuasion."). If that is shown, then the claim may be defeated. However, in a third step, if the employer has shown that a practice or policy is job-related and consistent with business necessity, a plaintiff can still prevail by demonstrating that there is "another selection device without a similar discriminatory effect that would also serve the employer's legitimate interest." Bradley, 443 F. Supp. 2d at 157. To reiterate, a claim of disparate impact discrimination does not require proof of an intent to discriminate. In fact, such a claim may be established even when the employer acts in good faith to avoid discrimination.

There is no "single test" to establish disparate impact. Langlois v. Abington Hous. Auth., 207 F.3d 43, 50 (1st Cir. 2000) (citing Watson v. Fort Worth Bank & Trust, 487 U.S. 977, 995-96 n.3 (1988) (plurality opinion)). Instead, "courts appear generally to have judged the 'significance' or 'substantiality' of numerical disparities on a case-by-case basis." Watson, 487 U.S. at 995 n.3.

One frequently used benchmark for identifying and measuring disparate impact is what is commonly referred to as the "four-fifths rule," articulated in the Uniform Guidelines on Employee Selection Procedures (1978) ("Uniform Guidelines") adopted by the Equal Employment Opportunity Commission ("EEOC"). 29 C.F.R. § 1607.4(D)[4] The four-fifths rule is not really a rule but a "rule of thumb," a rough guide suggested by the EEOC for assessing the

[4] EEOC v. Arabian Am. Oil Co., 499 U.S. 244, 257 (1991) (Because "Congress, in enacting Title VII, did not confer upon the EEOC authority to promulgate rules or regulations," the agency's guidelines receive weight only to the extent of their "power to persuade." (internal citations and quotations omitted)).

existence of disparate impact to be addressed by enforcement actions. See Watson, 487 U.S. at 995 n.3. According to the EEOC's Uniform Guidelines,

> A selection rate for any race, sex, or ethnic group which is less than four-fifths ($^4/_5$) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.

29 C.F.R. § 1607.4(D).

> The utility of the four-fifths rule may vary in different factual circumstances:
>
> Smaller differences in selection rate may nevertheless constitute adverse impact, where they are significant in both statistical and practical terms or where a user's actions have discouraged applicants disproportionately on grounds of race, sex, or ethnic group. Greater differences in selection rate may not constitute adverse impact where the differences are based on small numbers and are not statistically significant.

Id. Additionally,

> Where the user's evidence concerning the impact of a selection procedure indicates adverse impact but is based upon numbers which are too small to be reliable, evidence concerning the impact of the procedure over a longer period of time and/or evidence concerning the impact which the selection procedure had when used in the same manner in similar circumstances elsewhere may be considered in determining adverse impact.

Id.

The four-fifths rule was used by the parties and their experts. However, since the trial the First Circuit has counseled against too much reliance on the four-fifths rule. Jones v. City of Boston, 752 F.3d 38, 51 (1st Cir. 2014). For one thing, it may "lead to anomalous results." Id. It is not a precise measurement tool (it was not meant to be), and its principal utility in litigation may be to assist plaintiffs in making a sufficient prima facie demonstration of disparate impact. See Langlois, 207 F.3d at 50.

Disparate impact is itself a statistical construct; it is an inference of discriminatory practice from statistical evidence. See Ricci, 557 U.S. at 587 (describing a prima facie showing of disparate impact as "essentially a threshold showing of a significant statistical disparity . . . and nothing more"); Fudge v. City of Providence Fire Dep't, 766 F.2d 650, 658 (1st Cir. 1985) (a prima facie showing of disparate impact exists where "statistical tests sufficiently diminish chance as a likely explanation"). In Jones, the First Circuit recently emphasized the importance of preferring the analytic rigor of statistical analysis to the imprecise rule of thumb that the four-fifths rule supplies. 752 F.3d at 43, 52.

> Statisticians, by contrast, customarily approach data such as this more precisely. They ask whether the outcomes of an employment practice are correlated with a specific characteristic, such as race, and, if so, whether the correlation can reasonably be attributed to random chance. . . .
>
> To assess the likelihood that an observed difference in outcomes resulted from mere chance, statisticians calculate the probability of observing a difference equal to or greater than that which actually occurred, assuming equal opportunity. They call this probability the "p-value." Statisticians usually apply the label "statistically significant" to the observed differential outcomes if the p-value is less than five percent . . . .
>
> Essentially, a finding of statistical significance means that the data casts serious doubt on the assumption that the disparity was caused by chance.

Id. at 43 (internal citations omitted). The difficulty that plaintiffs often face in seeking to establish statistical significance is that large sample sizes are often required to show that the disparity is statistically significant. Id. at 53. As discussed further below, the plaintiffs here have that problem with every defendant employer except Boston.

The Uniform Guidelines also address when an employer's selection method is "job-related," such that its use might be justified notwithstanding some adverse impact on a particular group. Industrial psychologists generally describe a selection method that measures a candidate's knowledge, skills, and abilities (often expressed as KSAs) against the actual needs of the job as

being a "valid" selection tool. For an employer who seeks to justify a particular selection tool as "job-related," it is important to establish the tool's "validity" in this sense. Validity refers to the accuracy of inferences that an employer seeks to make about a candidate's suitability for the job on the basis of outcome, such as a test score, from a selection instrument. For example, an employer may administer an exam and hire individuals with the highest exam scores, the inference relied on being that the candidates with higher scores are more qualified than those with lower scores.

The Uniform Guidelines address the validity of selection instruments. See 29 C.F.R. § 1607.5. Industrial psychologists and the Guidelines recognize alternate possible measures of validity: criterion validity, content validity, and construct validity. Id. § 1607.5(A). The defendants in this case rely on content validity to justify use of the written civil service exams as a tool for selecting candidates for promotion from patrol officer to sergeant.

> Evidence of the validity of a test or other selection procedure by a content validity study should consist of data showing that the content of the selection procedure is representative of important aspects of performance on the job for which the candidates are to be evaluated.

Id. § 1607.5(B); see also id. § 1607.14(C). The other measures – criterion and construct validity – are not in issue.

The first step for content validation is the performance of a job analysis. Id. § 1607.14(C)(2). The next step is to ensure that there is a link between the selection procedure and the critical KSAs necessary for successful performance of the job as identified by the job analysis. Id. § 1607.14(C)(4). In addition, where a selection procedure relies on prior training or experience as a selection criterion, that criterion should be justified based on the relationship between the specific training, experience, and the job's requirements. Id. § 1607.14(C)(6). The use of a ranking device requires a separate demonstration that there is a relationship between

higher scores and better job performance. Id. § 1607.14(C)(9). A selection tool that validly assesses whether the candidate has KSAs necessary and appropriate to the job can be deemed job-related and consistent with business necessity for Title VII purposes.

As noted above, even if a selection tool is shown to be job-related and consistent with business necessity, it still may be condemned as discriminatory under the disparate impact theory if the plaintiff can show that there are other valid selection tools that generate less adverse impact on minority candidates. The inquiry then turns to specific alternate selection tools available to be used and the likelihood of reducing adverse impact by their use in place of the employer's chosen tool.

B. The Civil Service Regime and Written Job Knowledge Tests

The Court of Appeals summarized the Massachusetts civil service regime as it applies to Massachusetts police departments:

> Under Massachusetts law, plaintiffs' positions as city and MBTA police officers are subject to the state civil service law. See Mass. Gen. Laws ch. 31, § 48 (applying the civil service law to positions in the MBTA); id. § 51 (applying the civil service law to civil service offices in cities).
>
> The state civil service law states that the purpose of its requirements is to ensure that employees in civil service positions are recruited, chosen, and promoted based on principles of merit, not on political affiliation, race, age, gender, religion, national origin, or other factors unrelated to individual ability. Id. § 1. "[T]he fundamental purposes of the civil service system [are] to guard against political considerations, favoritism, and bias in governmental employment decisions . . . and to protect efficient public employees from political control." Cambridge v. Civil Serv. Comm'n, 43 Mass. App. Ct. 300, 682 N.E.2d 923 (1997).

Lopez, 588 F.3d at 74.

> By law, municipal police promotions must be made on the basis of competitive examinations, whether on the basis of the HRD examination or some other test. Mass. Gen. Laws ch. 31, §§ 59, 65. HRD is given statutory authority to establish the form and content of these examinations. Id. § 16. However, HRD's discretion in this area is bounded. By statute, all examinations must "fairly test the

> knowledge, skills and abilities which can be practically and reliably measured and which are actually required" to perform the job, a requirement that may significantly limit both the form and the substance of an examination. Id. And HRD must consult with labor representatives and professionals in the field to determine what skills and abilities are relevant for promotion to police sergeant or any other position. Id.

Id. at 77.

Massachusetts police and fire departments subject to civil service laws have two general options for promotional examinations. They may use statewide examinations developed by HRD, or they may seek approval from HRD to develop and use their own promotional exams. See Mass. Gen. Laws ch. 31 § 5(l) (giving HRD power to delegate administrative functions to cities and towns). For the years at issue, all defendants elected the first option, relying on HRD to design and administer the exams. Municipalities do not participate in the design or administration of the statewide exam in the absence of a delegation.

For each of the years in question, the HRD promotional exam for sergeants consisted of two elements: a written, closed-book exam consisting of 80 multiple-choice questions, and an "education and experience" ("E&E") rating. The E&E rating principally took account of relevant prior employment and academic coursework that a candidate had either taken or taught. The written exam accounted for 80% of the final score; the E&E component was assigned a 20% weight. Based on a 100-point scale, a candidate needed a score of 70 or above to pass the exam. In addition to the scoring of the exam, under Massachusetts law certain military veterans and long-service employees receive preference in the form of additional points that are added to their final exam score. Mass. Gen. Laws ch. 31, §§ 26 (veterans), 59 (long-service employees).

Participants in the test process are ranked by their combined score. HRD then prepares and certifies an "eligibility list" for each appointing jurisdiction, identifying those test-takers who may be considered for appointment to existing vacancies. Id. § 25. The number of names on

the eligibility list is determined by the formula 2n+1, where n is the number of vacant positions. Id. § 27. Thus, if a municipality had one job vacancy to which the list was applicable, the list would contain the candidates with the three highest scores (2x1+1=3). If there were three vacancies, the eligibility list would have the candidates with the top seven scores (2x3+1=7). In this case, it was the general practice of all employers except the MBTA to make selections in strict rank order according to the HRD eligibility list. The MBTA treated all candidates on the list as having scored equally and proceeded to make selections from that group principally on the basis of oral interviews of those candidates.

It should be noted that for statewide exams, HRD establishes an eligibility list for any municipality that requests one. Id. § 25. The list that is furnished to the municipality by HRD shows only those certified by HRD as eligible for appointment. The municipality does not receive information from HRD concerning other test-takers whose scores did not make them eligible for consideration under the 2n+1 rule. In other words, a municipality would not typically know which candidates did not score well enough to make the eligibility list, nor would the municipality know the race or ethnicity of test-takers who did not make the eligibility list. Consequently, the eligibility list alone would not ordinarily be a basis on which a municipality could determine whether there had been an adverse impact on minority test-takers or not.

Candidates may administratively appeal various issues pertaining to an examination, including "whether an examination . . . was a fair test of the applicant's fitness actually to perform the primary or dominant duties of the position for which the examination was held. . . ." Id. § 22. After administrative disposition of an appeal, further judicial review may be available. Id. § 44. Duties of the Massachusetts Civil Service Commission include adjudicating disputes concerning the content and administration of promotional examinations, any HRD decision or

action that affects an applicant, and any employment action taken by an appointing authority. Id. § 2(b)-(c). The Commission also "has the power to review any rules proposed by HRD, and, if the Commission concludes that a given rule violates a merit-based approach to employment decisions, it can, upon, a three-fifths vote, disapprove of the rule." Lopez, 588 F.3d at 75.

Under some circumstances, a municipal employer may skip over, or "bypass," a candidate on the list who would, by reason of his or her exam score, otherwise be selected for the vacancy. Id. § 27. But the employer must have a defensible reason for the bypass. Id. "If an appointing authority makes [a]…promotional appointment from a certification of any qualified person other than the qualified person whose names appear highest, and the person whose name is highest is willing to accept such appointment, the appointing authority shall immediately file with the [personnel] administrator a written statement of his reasons for appointing the person whose name was not highest." Id.; see also PAR 08(4), (5); PAR 09 (2).[5] An applicant who has been thus bypassed, i.e., not selected despite having a higher score than the selected applicant, can appeal to the Civil Service Commission, which must decide "whether the appointing authority has sustained its burden of proving that there was reasonable justification for the action taken by the appointing authority." City of Cambridge v. Civil Service Comm'n, 682 N.E.2d 923, 925 (Mass. App. Ct. 1997). A justifiable reason for a bypass might be, for example, a candidate's history of disciplinary infractions. Id. at 927 ("Prior misconduct has frequently been a ground for not hiring or retaining a police officer."). The candidate's race or ethnicity would

[5] PAR references are to HRD's Personnel Administration Rules. They "establish standards for the conduct of the civil service merit system of employment. In addition, these rules include standards governing state employment apart from civil service where rule making is required of the administrator by statute. They are intended to provide a system of uniform standards implementing applicable law for use by appointing authorities in the employment processes of recruitment and examination of applicants for public service positions, selection among applicants for appointment and promotion, performance evaluation and layoff." PAR 01.

ordinarily not be a justifiable reason for a bypass. Massachusetts Ass'n of Minority Law Enforcement Officers v. Abban, 748 N.E.2d 455, 461-62 (Mass. 2001) ("The commission, and the Superior Court judge on review, correctly concluded that without the consent decree's mandate, race, a consideration specifically identified by the Legislature in G.L. c. 31, § 1(e), as inconsistent with basic merit principles, cannot be used to justify a bypass.").

As noted above, municipalities may elect to design and conduct their own promotional examination pursuant to a delegation agreement between HRD and the municipality. See Mass. Gen. Laws ch. 31, §§ 9-11. However, even with respect to municipalities that have entered into a delegation agreement, HRD retains the right to approve the actions of the appointing authority. The appointment process after a local delegation is subject to the same rules regarding bypass and appeal or other challenge. It would also be subject to potential limitations imposed under collective bargaining agreements.

IV. Police Promotional Exams Litigation

The lower rates at which minority police officers in Massachusetts municipalities have been hired and promoted compared to non-minority officers has been a subject of considerable litigation over the past several decades. In 1970, black and Hispanic plaintiffs brought suit in this Court challenging hiring practices of the Boston Police Department ("BPD" or "Department"), including specifically the use of written examinations, as racially discriminatory and a denial of equal protection of the laws. After some litigation, the Court entered a consent decree requiring, among other things, that written examinations prepared and administered by the state Department of Personnel Administration ("DPA"), the predecessor to what is now HRD, be "validated in conformity with the Testing Guidelines of the Equal Employment Opportunity Commission, 29 C.F.R. §1607.1 et seq." Castro v. Beecher, 365 F. Supp. 655, 662 (D. Mass.

1973); see also Castro v. Beecher, 334 F. Supp. 930 (D. Mass. 1971), aff'd in part and rev'd in part, 459 F.2d 725 (1st Cir. 1972). Because the decree was addressed to DPA, it effectively applied to any municipality that used the statewide exam in hiring new police officers. The history of the Castro consent decree is summarized in Sullivan v. City of Springfield, 555 F. Supp. 2d 246, 248-50 (D. Mass. 2008).

In 1978, an association of black police officers filed suit against BPD and DPA alleging racially discriminatory practices affecting promotions from the rank of police officer to sergeant. See Massachusetts Ass'n of Afro-Am. Police, Inc. v. Boston Police Dept., 973 F.2d 18, 19 (1st Cir. 1992) (reciting history). In 1980 this Court again entered a consent decree, pursuant to which the defendants were limited to using promotional tests that were "specially validated as anti-discriminatory and fair." Stuart v. Roache, 951 F.2d 446, 448 (1st Cir. 1991). The consent decree was to expire in 1985, but because no validated promotional exams had been given by that time, the decree was extended for another five years to 1990.

Between 1985 and 1990, BPD administered a "validated-fair" exam, but promotions still fell short of the target goals set forth in the consent decree, and so the Court extended the decree again to permit an additional "validated-fair" exam to be given in an attempt to achieve the target goals. Id. An equal protection challenge by white police officers claiming to be disadvantaged by the continuing consent decree was rejected. See id. at 455 ("[T]he race-conscious relief here at issue represents a narrowly tailored effort, limited in time, to overcome the effects of past discrimination. As such, it is lawful. . . . And, the efforts in favor of eligible black police officers that it mandates therefore do not violate any statute of the Constitution of the United States."). The consent decree finally expired by its terms in 1995.

HRD administered another promotional exam for sergeants in 1996. In an effort to achieve more minority promotions, BPD bypassed some white officers on the certified list in order to promote three black officers who had scored one point lower on the exam. The white officers who were bypassed brought suit, claiming a violation of their equal protection rights. This Court rejected their claims, and the Court of Appeals affirmed, concluding that the promotions of three black officers constituted a narrowly tailored effort to overcome the continuing effects of past discrimination in hiring and was therefore not unlawful. Cotter v. City of Boston, 323 F.3d 160, 169-72 (1st Cir. 2003). A similar ruling also entered in a case involving a bypass promotion of a minority candidate to the rank of lieutenant based on a 1992 exam. See Boston Police Superior Officers Fed'n v. City of Boston, 147 F.3d 13, 19-25 (1st Cir. 1998).

Similar cases involving other police departments include Sullivan v. City of Springfield, 561 F.3d 7 (1st Cir. 2009) (City of Springfield) and Brackett v. Civil Serv. Comm'n, 850 N.E.2d 533 (Mass. 2006) (MBTA).

V. Disparate Impact of the Challenged HRD Exams for Sergeants

It is widely recognized among industrial organizational psychologists, including the four experts who testified at trial, that minority candidates as a group tend to perform less well relative to non-minority candidates as a group on written multiple-choice examinations such as the HRD-prepared sergeants promotional exams given in 2005, 2006, 2007, and 2008. Put in the language of disparate impact litigation, such exams are said to have an adverse impact on minority candidate pools, compared to non-minority pools. Commonly, adverse impact is assessed as the ratio of success rates of minority test-takers to the success rates of non-minority test-takers. Thus, as noted above, the EEOC has devised its "four-fifths" rule of thumb: "A selection rate for any race . . . which is less than four-fifths (4/5) (or eighty percent) of the rate

for the group with the highest rate will generally be regarded…as evidence of disparate impact." 29 C.F.R. § 1607.4(D).[6]

For purposes of assessing adverse impact, comparisons of success for different candidate pools can be done at various points. For example, adverse impact can be assessed at selection rates: what proportion of minority candidates are selected for hiring or promotion compared to the proportion of non-minorities selected. For example, suppose 10 minority candidates and 20 non-minority candidates compete for 5 available promotions, and 1 minority and 4 non-minority candidates are ultimately selected for promotion. The selection rate for the minority pool (1 of 10 or 10%) is compared with the selection rate for the non-minority pool (4 of 20 or 20%), producing an adverse impact ratio of .10/.20 or 50%. Comparisons can be done as appropriate at other potentially significant points, such as the passing score on the test, the effective passing score (at which one is eligible as a practical matter for hire or promotion), or mean group scores (how well groups as a whole perform, comparatively).

Reliance on statistical analysis to assess whether unlawful employment discrimination has occurred must be done cautiously, consistent with the limitations of the method, and there are circumstances where statistical methods are, by reason of their inherent limitations, unreliable tools for the purpose. One problem is the case of small statistical samples. With respect to small data sets, adverse impact ratios, standing alone, can be misleading. See Jones, 752 F.3d at 53; Fudge, 766 F.2d at 657-59.

For example, when a data set is small the shift of a single person from non-selection to selection for a job can alter an apparent conclusion regarding the presence of adverse impact. To illustrate, suppose of the thirty candidates in the example above one additional minority

[6] The reverse is not true. A selection tool that satisfies the four-fifths test is not necessarily non-discriminatory. Jones, 752 F.3d at 52.

candidate was promoted, so that two minority candidates and three non-minority candidates were promoted. The adverse impact ratio would shift from 0.50 to 1.33 (20% minority success rate / 15% non-minority success rate). The EEOC's guidance recognizes the problem of the "shift of one":

> If the numbers of persons and the difference in selection rates are so small that it is likely that the difference could have occurred by chance, the federal agencies will not assume the existence of adverse impact, in the absence of other information . . . Generally, it is inappropriate to require validity evidence or to take enforcement action where the number of persons and the difference in selection rates are so small that the selection of one different person for one job would shift the result from adverse impact against one group to a situation in which that group has a higher selection rate than the other group.

See Questions and Answers on the Federal Executive Agency Guidelines on Employee Selection Procedures, 44 Fed. Reg. 11, 996 (1979).

Importantly, proposed conclusions from small data sets can lack statistical significance. Assigning statistical significance is an attempt to ascertain whether apparent differences are the product of chance or of some other factor, such as a discriminatory selection method. In general, statisticians regard a selection result as statistically significant when the probability that the result is due to chance is less than 5%. Jones, 752 F.3d at 46-47, 47 n.9 (collecting cases where 5% is identified as statistically significant). It is common for small data sets to fail to produce results that are statistically significant.

Small data sets can also be statistically unstable. As defense expert Dr. Jacinto Silva explained, even where it is assumed that there are no performance differences between different groups, statistical analysis may yet indicate that there are by showing "false positive" adverse impact. (See Trial Ex. 197.) In other words, even on the assumption that there was no actual difference in the mean test scores between two groups, it can be a not uncommon consequence of the smallness of the sample sizes that adverse impact will appear. Consider an illustration from

Exhibit 197 using a municipality not a party to this suit. One minority officer and 18 non-minority officers of the Peabody Police Department took the 2006 sergeants exam. There were two promotions, both non-minority. On the assumption that there was no difference in performance on the test between the two groups, there was, according to Dr. Silva's analysis, an 89% chance that the data, conventionally assessed, would nonetheless suggest the presence of adverse impact. Suppose, for example, that among the 19 test-takers, the one minority candidate had the tenth highest score – that is, right in the middle of all the scores. Assuming for illustrative purposes a more or less even distribution of scores, the mean score averages for the one minority and 18 non-minorities would be exactly or very close to the same. Nonetheless, there would be a strong indication of adverse selection impact. That indication would be a product of the small numbers, however, and thus unreliable as a measure of actual adverse impact.

As Dr. Silva noted, while the EEOC has not directly addressed what might constitute a problematically small sample, it has used an example in which it characterized a sample of 100 as small. According to Dr. Silva, researchers have found that adverse impact ratios can be an unstable test with samples as large as 200 to 400 individuals. By these or any other similar standards, the sample sizes for each of the municipal defendants except Boston, including the MBTA, for the test years in question would qualify as "small," and would be subject to the cautions and limitations applicable to the statistical analysis of small populations.

In an attempt to overcome this problem of proof with respect to the smaller departments, the plaintiffs propose that test data for each relevant year should be aggregated for all municipalities across the Commonwealth that participated in the promotional exams for that year. So, for example, the plaintiffs argue that to assess whether the City of Methuen's use of the

HRD-sponsored test in 2006 had an adverse impact on minority candidates for promotion to sergeant in the Methuen Police Department that year, the Court should assess whether the HRD-sponsored test had an aggregate adverse impact statewide on promotions to sergeant, across all employing municipalities.

Aggregation of data across employing municipalities may be appropriate in some cases, but it is not appropriate in this case. Under Massachusetts law, a municipality may promote to sergeant only officers already employed in its own police department. Mass. Gen. Laws ch. 31, § 59. It may not select its sergeants from the ranks of police officers employed by other municipalities. As a result, a municipality's promotions should be assessed with respect to the pool of candidates actually available for appointment to the rank of sergeant. That pool consists of the candidates in its own department only. What adverse impact, if any, the test might have with respect to another municipality's candidate pool is simply not relevant.

Implicit in the plaintiff's aggregation argument is the assumption that there will be no difference by municipality in the composition of their respective pools of minority sergeant candidates, particularly with respect to vulnerability to adverse impact from the test. There is no reason to suppose, and no evidence on the point was produced at trial, that the assumption will necessarily prove to be true. There are a number of reasons why there may be substantial differences in the candidate pools in different municipalities. Since each pool of promotional candidates is comprised of persons hired by each municipality, variations in original selection procedures and subsequent in-service training, for example, could result in substantial variance in performance on the statewide exam between municipalities. Candidates in a department with strong training programs could be expected to be better prepared than candidates in a department with lax training programs.

Like small numbers, aggregation of pools of different sizes can produce anomalies, including one that statisticians refer to as Simpson's Paradox. Dr. Silva addressed this problem in Trial Exhibit 198. That exhibit presented a table showing the promotions made by two different hypothetical jurisdictions from their respective candidate pools. The pools were of different sizes. For each jurisdiction, the selection rate for minority and non-minority candidates was assumed to be identical. In other words, the assumption was that there was no adverse impact within either jurisdiction. When the data were aggregated, however, a false positive adverse impact was indicated, the result of combining different sample sizes and different selection rates. See Federal Judicial Center, Reference Manual on Scientific Evidence 233-35 (3d ed. 2011).

The plaintiffs also propose aggregating data for particular jurisdictions across exam years. For example, the City of Lawrence used the exam in 2006 and 2008. To ameliorate the problems arising from small sample size in each of the years, the plaintiffs propose assessing the combined results of those two test cycles. While that suggestion might have some superficial appeal, it has not been shown to be a reliable analysis technique. Forty-six Lawrence officers took the promotional exam in 2006, 10 minorities and 36 non-minorities. Forty-two took the exam in 2008, 15 minorities and 27 non-minorities. The numbers cannot simply be added together to conclude that there were 88 separate test-takers of whom 25 were minority and 63 were non-minority. Rather, it is very likely that there would be overlap in the various pools, since it is not uncommon for officers to take tests more than once. What the effect would be of the same test-takers in both years would necessarily be a matter for speculation. One test taker might be simply not qualified for promotion; his presence in the pool in both years might exaggerate the likelihood of failure for an aggregated pool. Another test taker might have benefitted from a lack of success the first time and with familiarity with the process and perhaps

heightened preparation and thus succeeded in the second process. There is simply no way to adjust for the possible variations that could distort interpretation of the aggregated data.

The plaintiffs' expert, Dr. Joel Wiesen, relied on aggregated data, both over time and across jurisdictions, for most of his essential conclusions about the adverse impact of the HRD-sponsored exams. Because I reject the use of aggregated data, I find his conclusions unpersuasive.

With this background, on the basis of the evidence adduced at trial, I make the following findings regarding the existence of adverse impact in promotions to sergeant by the various defendants:

The use of the HRD-sponsored sergeant's promotional exam by the City of Boston in 2005 and 2008 in each year had a significant adverse impact on black and Hispanic test-takers. Boston does not contest, indeed concedes, this finding. It rests its defense in the case on its contention that the exam was nonetheless job-related and consistent with business necessity and that the plaintiffs are unable to demonstrate an adequate alternative that could be used with less adverse impact. Those matters are addressed further below.

On the other hand, as to each of the other municipal defendants and the MBTA, the statistical evidence relied on by the plaintiffs is not sufficient by itself – and it stands essentially by itself – to persuasively establish that the use of the HRD-sponsored exam by those employers was the cause of an adverse impact on minority promotion rates in those various jurisdictions. The principal problem is the small sizes of the relevant data sets, for the reasons discussed above.

The plaintiffs had relied in their claims of adverse impact on the proposition that statistical data should be aggregated across jurisdictions. I have rejected that approach, as discussed above. Without aggregation, the plaintiffs' statistical case against the jurisdictions other than Boston falls apart.

A. Worcester

Plaintiff Spencer Tatum, an African-American, participated in HRD-sponsored sergeants examinations for the Worcester Police Department in 2006 and 2008. The following table reflects the results of these exams.[7]

Worcester		
Year	2006	2008[8]
Minority Test-Takers	10[9]	8
Minority Appointments	1	0
Minority Appointment Rate	9%	0%
Non-Minority Test-Takers	51	47
Non-Minority Appointments	6	1
Non-Minority Appointment Rate	12%	2%
Adverse Impact Ratio	0.90	0

For 2006, the ratio of minority appointments to non-minority appointments is .90, suggesting under the four-fifths rule (though not proving) the absence of adverse impact. As of the time of trial, only one appointment had been made from the 2008 exam, and an adverse impact ratio is not calculable.

In any event, in light of the small numbers involved, the difference in appointment rates between minorities and non-minorities in Worcester for both the 2006 exam and the 2008 exam

[7] The data in this and the following sections are generally derived from Trial Exhibit 197.

[8] The data for 2008 are from Trial Exhibit 175.

[9] In Exhibit 197, Dr. Silva had reported that in 2006 there were 11 minority test-takers and 50 non-minority test-takers. At trial, Dr. Silva corrected the data to 10 minorities and 51 non-minorities. The error was apparently due to an incorrect self-report by one of the test-takers. (Trial Tr., day 17, at 20.)

are not statistically significant. According to Dr. Silva, the Fisher's Exact p-value for the 2006 exam is rounded to 1.0, suggesting that the difference in appointment rate may well be the result of random chance.

B. Springfield

Plaintiffs James A. Jackson, Juan Rosario, Louis Rosario, Jr., Obed Almeyda, Devon Williams, and Julio M. Toledo participated in HRD-sponsored sergeants exams in 2005 and 2007 for the Springfield Police Department.[10]

The 2005 and 2007 exams resulted in the following data for Springfield:

Springfield		
Year	2005	2007
Minority Test-Takers	18	16
Minority Appointments	0	2
Minority Appointment Rate	0%	12.5%
Non-Minority Test-Takers	28	20
Non-Minority Appointments	6	6
Non-Minority Appointment Rate	21%	30%
Adverse Impact Ratio	0	0.42

While both the 2005 and 2007 exams indicated adverse impact as to the appointment rate under the four-fifths rule, the results were not statistically significant. The p-value for the 2005 exam was .07 and for 2007 the p-value was .26. A p-value of .05 is commonly used by social scientists as necessary to reject the "null hypothesis." Since the statistics for both years do not meet that standard, the null hypothesis – here, that there is no adverse impact – cannot be rejected on statistical calculation alone. Additionally, as Exhibit 197 indicates, for both years

[10] It appears that the Springfield plaintiffs did not timely file a necessary pre-suit claim with the Massachusetts Commission Against Discrimination with respect to the eligibility list from the 2005 exam, and therefore have no viable claims regarding that exam.

there was a substantial possibility of a false positive adverse impact finding. As Dr. Silva explained, the high false positive estimate results from

> a combination . . . of the small sample size, the small number of minorities, the selection ratio. Those three things contribute to the false positive rate. . . . With a small sample size, the data is always unstable.

(Tr. 17: 44-45.)

In sum, the statistical evidence is unconvincing as to the existence of adverse impact.

It may also be significant that in the relevant years promotions to sergeant were made using the eligibility list generated from the exams, supplemented by an interview process and a review of the candidates' department work history. The evidence does not show whether or how the interviews and/or work history of applicants may have affected appointments.

C. Lowell

Plaintiff Robert Alvarez participated in the HRD-sponsored sergeants exam in 2006 for the Lowell Police Department.

The Lowell 2006 sergeants exam resulted in the following appointment data.

Lowell	
Year	2006
Minority Test-Takers	7
Minority Appointments	0
Minority Appointment Rate	0%
Non-Minority Test-Takers	36
Non-Minority Appointments	7
Non-Minority Appointment Rate	19%
Adverse Impact Ratio	0

Again, because of the small numbers, the statistical evidence is not reliable to show that the use of the HRD-sponsored exam caused an adverse impact. The insufficiency of the data to support any reliable statistical conclusions is demonstrated by a shift-of-one analysis. If one minority test taker (instead of none) and 6 non-minority test-takers (instead of 7) were appointed,

then the adverse impact ration would be .86, suggesting no adverse impact under the four-fifths rule. Moreover, the p-value for the plaintiff's proffered adverse impact ratio (per Dr. Wiesen) in the selection rate for promotions 2006 exam is .58, indicating a lack of statistical significance.

Status of Alvarez as an Aggrieved Plaintiff

Lowell also challenged Alvarez's standing to claim discrimination on the ground that he has not shown himself to be "Hispanic" and thus a "minority" entitled to complain about the sergeants exam's discriminatory impact on such minority officers. Although his claim would fail for the reasons just discussed, because the parties thoroughly addressed his status at trial, it is appropriate to resolve this issue.

The evidence indicated that Alvarez was born in Boston and was raised by a foster family in Boston-area suburbs, including Somerville and Billerica. (Tr. 9: 118-19, 146; Trial Ex. 136.) On his birth certificate, his birth mother and father are described as white. (Tr. 9: 126-27; Trial Ex. 136.) His birth certificate also indicates that his father, whom Alvarez apparently never met and with whom he has never spoken, was born in Manila, Philippines. (Tr. 9: 115-16; Trial Ex. 136.)

Alvarez does not speak Spanish. (Tr. 9: 113.) He was not raised in a home where Spanish was the primary language. (Id.) At times he has identified himself as white, even when Hispanic was an identified option (Tr. 9: 123-27; Trial Exs. 138, 139) and at other times he has self-identified as Hispanic. (Tr. 9: 103.)

An unambiguous definition of "Hispanic" for purposes of anti-discrimination laws has long eluded both courts and legislators. For instance, for record-keeping purposes, EEOC has used "Hispanic" in reference to "persons of Mexican, Puerto Rican, Cuban, Central or South American, or other Spanish culture or origin, regardless of race." See 29 C.F.R. § 1607.4. This,

by itself, is not helpful in resolving Alvarez's claim to be considered Hispanic on the basis of Filipino ancestry. It is of interest that the EEOC requires that employers report employees with origins from the "original peoples" of the Philippines as "Asian or Pacific Islander." EEOC Instruction Booklet for EEO-1 Report (2006). The trial record includes no evidence whether Alvarez's father had ancestors who were properly considered among the "original peoples" of the Philippines.

Definitions used by HRD provide a bit more guidance. HRD defines "Hispanic" as an "individual who (or whose family) originates from a Spanish-speaking country in the Western Hemisphere and who either speaks Spanish or was raised in a household where Spanish was the primary language." (Aff. of Sally McNeely ¶ 5 (Ex. 205) (dkt. no. 307).)[11] HRD categorizes an individual who (or whose family) originates from the Philippines as Asian, not Hispanic. (Id. ¶ 6.) Under HRD's definition, Alvarez, whose father originated in the Philippines and who does not speak Spanish and was not raised in a household where Spanish was the primary language, would not be classified as Hispanic for Civil Service purposes.

In this case, it seems appropriate to defer to HRD's understanding of the term "Hispanic." Applying that understanding, Alvarez does not qualify, and his claim would fail for this additional reason.

D. Lawrence

Plaintiffs Pedro J. Lopez, Richard Brooks, and Kevin Sledge took the sergeants exam in both 2006 and 2008 for the Lawrence Police Department. The results of that exam can be summarized as follows:

[11] The definition apparently derives from two federal consent decrees arising out of unrelated litigation involving allegations of discrimination.

Lawrence		
Year	2006	2008
Minority Test-Takers	10	15
Minority Appointments	0	0
Minority Appointment Rate	0%	0%
Non-Minority Test-Takers	36	27
Non-Minority Appointments	3	1
Non-Minority Appointment Rate	8%	4%
Adverse Impact Ratio	0	0

Again, the small data sets prevent any reliable conclusion from statistical analysis alone. A shift-of-one analysis illustrates this. If in 2006 there was one minority appointment (as opposed to none) and two non-minority appointments (as opposed to three), the resulting adverse impact ratio (2.0) would satisfy the four-fifths rule of thumb. Additionally, of course, because of the small numbers, any adverse impact ratios for either year would not be statistically significant.

E. Methuen

Plaintiff Abel Cano, who is Hispanic, participated in the HRD-sponsored sergeants exam in 2006 and 2008 for the Methuen Police Department. The following data summarize results from those exams.

Methuen		
Year	2006	2008
Minority Test-Takers	4	3
Minority Appointments	0	0
Minority Appointment Rate	0%	0%
Non-Minority Test-Takers	19	15
Non-Minority Appointments	1	0
Non-Minority Appointment Rate	5%	0%
Adverse Impact Ratio	0	

There are no statistically reliable indicators that show minority applicants in Methuen were adversely affected by either the 2006 or 2008 sergeants promotional exams. The results of the adverse impact analysis by the plaintiffs' expert of the selection rates in Methuen in both the 2006 and 2008 sergeants promotional examinations were not statistically significant.

F. MBTA

Plaintiffs Royline Lamb and Lynn Davis, who are both African-American, participated in the HRD-sponsored sergeants exam in 2005 for the MBTA Police Department, and Lamb participated also in the 2007 exam. Davis did not receive a passing score on the 2005 exam. She was informed that she did not pass the exam when the results were published by HRD several months after she took the exam. She did not take the 2007 HRD exam for promotion to sergeant. Lamb did not receive a passing score on either exam. Like Davis, he learned of his failure to pass the 2005 exam several months after its administration.

On September 24, 2008, Davis and Lamb filed charges of discrimination with the Massachusetts Commission Against Discrimination ("MCAD") pursuant to Mass Gen. Laws, ch. 151B, §4. The applicable statute of limitations for filing a charge of discrimination at the MCAD is 300 days from the discriminatory act. Mass. Gen. Laws ch. l51B, §5. Davis did not take the 2007 HRD exam for promotion to sergeant so her claim, relating only to the 2005 exam, was untimely under applicable statute of limitations, and it must be dismissed. Allston v. Massachusetts, 661 F. Supp. 2d 117, 123 (D. Mass. 2009). For similar reasons, Lamb's claim regarding the 2005 exam is likewise time-barred.

Limitations issues aside, the existence of disparate impact has not been proven. Since neither Davis nor Lamb passed the exam, the focus for examining possible adverse impact is on the relative rates at which minorities and non-minorities passed the exam. The adverse impact ratio for passing rates for 2005, as computed by Dr. Wiesen, was .32. For 2007, it was .96. (See Trial Exs. 83, 197.) The p-value for the 2005 test was 0.31, well outside the range of statistical significance. Not surprisingly, for 2007 the p-value was 1.00.

In summary, for all jurisdictions except Boston, the plaintiffs have not carried their burden of proving a prima facie case of disparate impact from any of the challenged exams.

VI. Job-Relatedness / Business Necessity

As discussed above, only for Boston have the relevant plaintiffs shown sufficient adverse impact on minority promotions to warrant an inference of disparate impact discrimination. Under governing principles, the next question is whether Boston has demonstrated that the civil service examination nonetheless was job-related and consistent with business necessity, and thus lawful. See 42 U.S.C. § 2000e-2(k)(1)(A)(i). Boston bears the burden of proof as to this issue. Id.

As previously discussed, industrial organizational psychologists refer to a selection instrument that is job-related and consistent with business necessity as one that is "valid" for the selection process. Dr. James Outtz, a prominent industrial psychologist and Boston's expert witness, explained the concept of validity as it is understood in the field of industrial psychology:

> In my field, from a scientific standpoint, validity refers to the accuracy of inferences that you wish to make on the basis of scores from a selection instrument: Are those inferences accurate? For example, if you give an exam and you hire people at the – who score the highest on that exam, the inference is that those people are more qualified than people who scored lower. If you give an exam for people to get admitted into college, and you admit students whose scores are up at the upper distribution, you are making the inference they are more qualified than students who have lower scores. Validity refers to evidence that is garnered to establish the accuracy of those inferences. In a more simplistic way, it's whether a test measures what it's supposed to measure.
>
> That is the essence of validity, to allow one to predict at a level great than – significantly greater than chance as to how well someone will perform.

(Tr. 14: 15-16.)

Massachusetts Civil Service law requires promotional appointments within police departments, including the Boston police department, to be made after competitive examination. Mass. Gen. Laws ch. 31, § 59. Such examinations are ordinarily prepared by HRD. Id. § 16.

Specifically, the examinations used by Boston in 2005 and 2008 were prepared by HRD. Such examinations must "fairly test the knowledge, skills and abilities which can be practically and reliably measured and which are actually required to perform the primary or dominant duties of the position for which the examination is held." Id. As noted, an examination that does reliably test those KSAs is deemed to be "valid" for the purpose of identifying the candidates best suited for selection.

In order to create a valid examination, it is necessary first to conduct a job analysis to determine what the job actually entails. The job analysis seeks to identify the tasks that make up the job and the KSAs required to perform these tasks. This is typically accomplished by consultation with persons familiar with the position in question, generally referred to as "subject matter experts," or SMEs.

Once the job analysis has been done, it is necessary to develop a test that will measure the necessary KSAs. While a test plan should measure as many of the pertinent KSAs as practical, under the EEOC's Uniform Guidelines on Employee Selection Procedures, see 29 CFR §§1607.14(C)(4) and 1607.15(C)(4) (pertaining to content validity), a testing instrument need only assess a representative sample of the KSAs required for the job. Once a test plan is developed, it is necessary to develop the actually content of the examination and to have it reviewed by SMEs.

The continuing validity over time of a job analysis and any test instrument developed in reliance on it depends on whether and to what degree the demands of the job may change over time. As a rule of thumb, industrial psychologists regard a job analysis performed within five to eight years of an examination to be reliable, but sometimes an older one can still be appropriate if the requirements of the job itself have remained more or less the same. Additionally, an older job analysis might be reevaluated and updated in a later review.

The position of sergeant in the Boston police department is a supervisory one. Sergeants are typically field supervisors for patrol officers. It is essential that sergeants be familiar with constitutional and statutory legal principles pertinent to their jurisdiction, as well as departmental regulations and policies. The Boston Police Commissioner and a former Boston Police lieutenant both testified at trial that it is critical to a police sergeant's ability to effectively perform as a supervisor that he or she know and understand relevant law. When patrol officers need information or clarification, the first thing they do is to call their sergeant. An effective way of testing whether a candidate for sergeant has the necessary knowledge is through a written job knowledge test.

The 2005 and 2008 written exams prepared by HRD relied in substantial part on a content validation study for a promotional exam for sergeant that was conducted in 1991 by DPA, HRD's predecessor. In preparing the Validation Report, DPA first "conducted a comprehensive job analysis of superior officer ranks, including…sergeant…" (Trial Ex. 40, Bates Stamp page 247). A principal feature of this analysis was "[g]athering of available job information from Massachusetts police departments, as well as job analysis reports, survey instruments and other information from jurisdictions outside the Commonwealth." (Id. at 249).

In addition, DPA "[d]eveloped and administered a task inventory questionnaire designed to identify the frequent and critical tasks and duties," and "a knowledge, skills, abilities and personnel characteristics (KSAPs) inventory questionnaire designed to identify the important KSAPs required at the time of appointment." (Trial Ex. 40, Bates Stamp page 249). DPA also developed a "[l]inkage of the important KSAPs to the frequent and critical tasks of these jobs from" SMEs, "designed and use[d]…structural group discussions to gather information from SMEs about the Education and Experience (E&E) component of DPA's selection procedures,"

and "gather[ed] information from SMEs about the recommended reading list from which the multiple choice written examination questions for the police promotional exams are derived." (Id. at 250).

The preparation of the Validation Report also included work by a Job Analysis Team ("JAT"). The JAT requested and received job descriptions for sergeant from Massachusetts police departments covered by Civil Service. The JAT also submitted a task inventory questionnaire for sergeant, covering 136 task or duty statements, to a sampling of persons at 76 police departments in Massachusetts. The JAT received 824 responses, a response rate of 78%. DPA used the responses to develop task profiles for sergeants. Specifically, it developed a list of KSAs that it distributed to SMEs for their review and comment, who were generally superior officers serving in Massachusetts police departments. (Trial Ex. 40, at Bates Stamp page 257). The JAT also convened a group of nine SMEs from Massachusetts police departments to review a master list of tasks that had been identified by the general survey as important and frequently performed.

Exams to be developed after the job analysis was completed were also to include a new E&E component pursuant to which incumbents would receive points for past educational achievements and work experience that in combination with their raw score on the written civil service exam become the score by which they will be ranked and placed on civil service eligibility lists.

The JAT designed a structured discussion guide to identify what educational degrees, certificates, or licenses were important for sergeants under consideration, as well as relevant prior work experience. The JAT arranged for SMEs from various Massachusetts police

departments to participate in discussion groups to obtain feedback regarding the E&E component.

A "testability analysis" identified KSAs that could be assessed under both the written exam and the E&E. (Attachment EE to the 1991 Validation Report, Trial Ex. 41, at Bates Stamp pages 4272-4278). The testability analysis showed that more than half of the KSAs identified as pertinent to the job of sergeant were tested. This was sufficient to meet the "representative sample" requirement of the Uniform Guidelines. (Tr. 14: 41-43.)

In addition to performing a job analysis, the JAT prepared a reading list that could be used by candidates to prepare for the exam to be given. The job knowledge questions on the exam would be based on information directly presented by materials on the reading list. The purpose was to permit a candidate to study effectively for the eventual exam by reading the materials on the list. The list was compiled based on input from SMEs.

In 2000, a consulting firm, Morris & McDaniel, performed a job analysis for the position of sergeant in the Boston Police Department, and that analysis was used in the development of Boston's subsequent 2002 sergeant exam. The job analysis was not as full a validation study as was done in 1991, but it generally supported the use of a written job knowledge test for a sergeants promotional exam.

Consistent with its prior practice in preparing other police promotional exams for Boston, the exams prepared by HRD for the position of BPD sergeant in 2005 and 2008 were based on a test plan that can be traced to the 2000 job analysis prepared by Morris & McDaniel, which itself can be traced to the 1991 job analysis, including particularly the KSAs identified in those documents.

In the regular process of developing the 2005 exam, HRD consulted with a panel of BPD captains (i.e., SMEs), some of whom were deputy superintendents, who recommended a reading list that had been periodically updated over time, but was consistent with the 1991 job analysis. The 2005 BPD promotional exam included certain questions specific to BPD, and as a result, the reading list included BPD rules, procedures, and special orders.

The 2005 sergeants promotional exam for Boston tested KSAs identified throughout the 1991 and 2000 job analyses related to the following subject matter areas: questions 5, 6, 8, 10, 57-60, and 62-67 addressed police management issues; questions 9, 11, 14, 16, 23, and 30 addressed crime-related issues; questions 19, 20, and 26 addressed crime scene issues; questions 21, 47, 49, and 50 addressed interrogation; questions 22, 24, and 25 addressed child abduction; question 27 addressed hostages; questions 41 through 46 addressed arrest procedures; questions 51 through 54 and 56 addressed searches.

The results of the examination for promotion to lieutenant provide some further evidence of validity of the 2005 sergeants exam. The 2005 written exams for the sergeant and lieutenant positions at BPD in 2005 had 53 items in common. Candidates for promotion to lieutenant, who were necessarily all then-incumbent sergeants, had an 89% passing rate on these common questions, whereas the passing rate for patrol officers seeking promotion to sergeant on the same questions was only 63%. The high passing rate for incumbent sergeants on the common questions is evidence that those questions were related to the sergeants' actual performance of their jobs. In other words, the sergeants scored well on those questions because the questions addressed issues related to their actual work experience. This fact tends to contradict any argument that the 1991 job analysis on which the test plan was partially based was dated and outmoded.

In the promotional process in 2008, Boston again utilized the services of HRD to prepare a sergeants promotional exam. This exam, like the 2005 exam, included BPD-specific questions. In preparation for this exam, HRD again prepared a test plan showing areas of knowledge to be tested and the number of items to be devoted to each area.

As with the 2005 exam, SMEs updated the reading list for candidates. Documents from HRD show that for the 2008 sergeant promotional exam for Boston, SMEs reviewed test items, indicated which KSA the item was linked to, assessed the difficulty level of the item, and recommended whether or not the item should be used. As with the 2005 exam, the 2008 sergeants promotional exam for Boston contained a number of questions focusing on specific rules, regulations and special orders of BPD.

The 2008 sergeants promotional exam for Boston tested KSAs identified throughout the 1991 and 2000 job analyses and related documents. (See Trial Ex. 45.) For example, questions 21 and 36 addressed juvenile issues; questions 2, 3, 33, 42, 47, 50, and 58 addressed various crime-related issues; questions 4 through 6, 22, 40, 41, and 51 addressed searches; questions 7 and 8 addressed firearm issues; questions 23 through 30, 34, 48, 49, 53, and 70 through 73 addressed police management issues; questions 9 through 16 addressed rules and regulations of BPD; questions 31, 32, 35, 37, 74, and 75 addressed community policing; questions 43 and 52 addressed interrogation; questions 54 through 57 addressed illegal drugs; and questions 76 through 80 addressed the ability to read and understand reading charts.

In an effort to assist candidates who would be taking promotional exams, in both 2005 and 2008 BPD engaged an outside consultant to provide extensive tutoring for any interested candidate. Prior to the 2005 promotional exam, BPD offered tutorials and study materials for

candidates at no cost. Prior to the 2008 exam, BPD similarly provided compact discs containing lectures prepared by the consultant for use by officers on their own time.

For both the 2005 and 2008 promotional exams for Boston, candidates for police sergeant completed an "Education and Experience Rating Sheet" in accordance with instructions thereto. Section III of the E&E Rating Sheet asked candidates to provide information on their police experience. As a matter of statutory law, police officers must have three years' experience to be eligible to apply for promotion to sergeant. Mass. Gen. Laws ch. 31, § 59. In addition, seniority – based on the number of years on the job – is generally recognized as relevant to the ability to perform well in a supervisory position such as sergeant.

Section IV of the E&E Rating Sheet asked candidates to indicate whether they had earned various academic degrees or certifications in various specified subject areas. As part of the 1991 Validation Report, SMEs had agreed that these subject areas were related to the position of sergeant.

Section V of the E&E Rating Sheet requests candidates to indicate if they had taught any courses above the high school level in the same subject areas for which credit is given if a degree is earned in that area. The SMEs who participated in the 1991 Validation Report concluded that the ability to teach a course evidenced oral communication skills important to the position of sergeant.

Dr. Outtz, the industrial psychologist, opined at trial that the tests as administered for Boston in 2005 and 2008 were "minimally valid." He acknowledged the utility of a written job knowledge test in selecting candidates for promotion to sergeant, but noted that use of such a test by itself would not support a conclusion of validity, because it could not measure some skills and abilities (as distinguished from knowledge) essential to the position, such as leadership, decision-

making, interpersonal relations, and the like. However, because he thought that such skills and abilities were attested to by the E&E component, the threshold of validity was crossed.

After consideration of the evidence as a whole, I find and conclude that Dr. Outtz's opinion rests on adequate grounds and is therefore correct: the exams in question were minimally valid. The exams satisfied the technical standards for content validity studies. 29 C.F.R. § 1607.14(C). They addressed a representative sample of the KSAs of the sergeant position. Id. § 1607.14(C)(1), (4). They were based on job analyses that considered the important tasks necessary to the successful performance of the job. Id. § 1607.14(C)(2). They took account of prior relevant work experience as well as relevant training and education. Id. § 1607.14(C)(6).

There is no doubt that Dr. Outtz also thought that the content validity of the exams could have been improved by the use of additional test elements, such as an assessment center (see infra). However, for assessing validity, the fact that it could have been better does not mean necessarily that it was not good enough to be deemed sufficient. The key question regarding the content validity of a selection method is whether it reliably predicts a candidate's suitability for the job, such that persons who perform better under the test method are likely to perform better on the job. I am satisfied on the evidence that Boston carried its burden of showing that the exams in question satisfied that criterion.

A selection method that is valid under the considerations discussed above can be deemed to be "job related" and "consistent with business necessity" under the statutory standard. 42 U.S.C. § 2000e-2(k)(1)(A)(i). I find that the City of Boston has successfully demonstrated by the evidence that that standard has been met.

VII. Availability of an Equally Valid, Less Discriminatory Alternative

A. General principles

If an employer has established that its selection method is "job related for the position in question and consistent with business necessity," as I have concluded Boston has done here, the plaintiffs may nonetheless succeed in their claims of disparate impact discrimination if they are able to establish that there was an "alternative employment practice" that Boston refused to adopt that was equally valid and would have had a lesser discriminatory impact. 42 U.S.C. § 2000e-2(k)(1)(A)(ii), (C); Albemarle Paper Co. v. Moody, 422 U.S. 405, 425 (1975); see also Ricci, 557 U.S. at 587.

The expert witnesses for both sides in this case agreed on two propositions about the use of written job knowledge tests in making promotions from police officer to sergeant. First, they agreed that properly developed written examinations are generally valid for that purpose. Second, they agreed that, as a general matter, the scores achieved by black and Hispanic test-takers on such written examinations in the aggregate tend to be consistently lower than the scores achieved by non-minority test-takers in the aggregate. In other words, experience has demonstrated that in the aggregate the use of written exams alone tends to have an adverse impact on minority applicants for promotion. Consequently, one of the major projects for industrial psychologists in the last several decades has been to try to develop selection methods that have the same validity as written job knowledge tests – that is, methods that select the better qualified applicants for promotion over the lesser qualified – but have a lesser or no adverse impact on minority applicants in the aggregate.

Commonly, the approach has been to supplement written job knowledge tests by adding other components to the overall selection method. A number of different options can be used to provide such supplementation. Some options that have been used include so-called "assessment centers," which may use oral interviews, role-playing exercises, writing samples, in-basket exercises, and group exercises. Explicit consideration of prior work performance is another option.[12]

Some of the municipal defendants in this case, including Boston, have utilized a number of these methods to supplement a written exam for certain hiring or promotion decisions. For example, for the purpose of selecting a chief of police in the 1990s, Lowell used a structured interview, a mock press conference, an in-basket exercise, and an essay. Springfield and MBTA interview candidates who are on the certified Civil Service list, as well as considering prior work. Boston's prior use of an assessment center is discussed below.

Assessment centers are generally used only as a supplement to, rather than a substitute for, written job knowledge exams for a couple of reasons. In the first place, civil service regimes typically require public employers to use written exams as a merit selection tool. That is true in Massachusetts for promotion within municipal police departments. See Mass. Gen. Laws ch. 31, § 59. Moreover, for many of the KSAs required for the job, written job knowledge tests are highly valid and thus preferred to other possible ways of assessing those KSAs. There are some practical reasons as well. Assessment center exercises can require considerably more resources to administer, including both money and personnel, and thus can be cumbersome and expensive. As a result, the practicality of their use tends to be inversely proportionate to the number of job candidates to be assessed. For example, it is easier and much less costly to have a multi-

[12] With respect to the 2002 BPD sergeants exam, consideration of prior work performance was prohibited by the applicable collective bargaining agreement. (See Trial Ex. 133 at pg. 41.)

component selection method for chief of police where there may be a relatively small pool of candidates than it would be for the rank of sergeant where, in Boston's case, the pool of interested candidates typically numbers several hundred. That is why some jurisdictions, such as Springfield and MBTA, use assessment centers as a second step after an original pool of test-takers has been narrowed by grade ranking.

B. Boston's Testing History

1. 1973-1998

Prior litigation concerning claims of discrimination in hiring and promotions in the Boston Police Department has been summarized above. Throughout most of that time, and certainly from the 1980s on, the Department has pursued the goal of reducing or eliminating disparate impact in its hiring and promotional selection methods and thus increasing the number of minorities in all ranks, including sergeants. Initially, under the consent decrees, the approach had been to seek a remedy in affirmative promotion formulas. Thus, with regard to promotions to sergeant and other supervisory ranks in BPD, this Court entered a consent decree in 1979. The decree "contained various affirmative action provisions designed to increase the number of black officers promoted to sergeants." Massachusetts Ass'n. of Afro-Am. Police, Inc. v. Boston Police Dept., 780 F.2d 5, 6 (1st Cir. 1985). Under the MAAAP decree, BPD was authorized to select individual African-American candidates for sergeant out of strict rank order as determined by results from an examination process, up to a potential 20% of all promotions.

As the legal landscape shifted in more recent years, however, the emphasis turned to fine-tuning selection methods with an eye toward reducing or eliminating adverse impact of the testing methods so that the cohorts from which selections would be made would not be discriminatorily constituted. In 1987, BPD, through a delegation agreement with DPA, retained

Morris & McDaniel to prepare a sergeants and lieutenants promotional exam that would include a multiple choice written examination, an in-basket exercise, a video performance exercise, a leaderless group exercise, and an education and experience component. The in-basket, video performance, and leaderless group exercises constituted the assessment center. Unfortunately, the integrity of the assessment center component was compromised by the leak of information about its contents before the examination, and the Personnel Administrator of DPA concluded that the assessment center scores could not be used in making promotional decisions. As a result, only the multiple choice exam and the education and experience components were used.

After extension of the MAAAP decree, BPD administered a "validated-fair" exam in June 1991. After the administration of the 1991 exam, the Massachusetts Association of Minority Law Enforcement Officers ("MAMLEO") (MAAAP's successor organization) challenged the exam's validity on the ground that it did not comport with the Equal Employment Opportunity Commission's Uniform Guidelines. See Boston Police Superior Officers Fed., 147 F.3d at 15-16. MAMLEO's legal challenge was settled by means of an amendment to the MAAAP consent decree, which provided, among other things, "that the BPD would establish the next eligibility lists for promotion to sergeant…using selection procedures 'of a significantly different type and/or scope.'" Id. at 16.

In 1992, a sergeants exam was administered which consisted of a written exam and a presentation by candidates to a group of assessors, consisting of police commanders from throughout New England. Boston's costs for the 1992 exam were in the range of $500,000. After the administration of the 1992 exam, Boston bypassed some white candidates on the certified eligibility list to reach and promote some black officers who had slightly lower scores. The white officers filed a complaint with the Civil Service Commission challenging the bypasses. The

Commission agreed that the bypasses were not justified and ruled against the promotion of the black officers. The Commission's order was ultimately upheld by the Massachusetts Supreme Judicial Court. See Massachusetts Ass'n. of Minority Law Enforcement Officers v. Abban, 748 N.E.2d 455 (Mass. 2001). Consequently, despite the inclusion of an assessment center in the selection process, there was no improvement in minority promotions. In 1998, BPD participated in the statewide promotional exam for sergeants prepared and administered by HRD. The exam did not include an assessment center.

2. 2002 Promotional Exam

After the 1998 exam, then Police Commissioner Paul Evans convened a committee in an effort to increase minority representation in the promotional ranks. Based on the committee's recommendation, the Department decided to include an assessment center in its next promotional process for the ranks of sergeant, lieutenant, and captain, along with a written exam, and it obtained an authorized delegation from HRD to do so. Under the delegation, BPD was responsible for funding and administering the exam, subject to HRD's oversight. BPD again retained Morris & McDaniel to prepare and administer the exam. It was in this connection that Morris & McDaniel performed the 2000 job analysis of the position of sergeant at BPD discussed in the validity section above.

Under the delegation agreement, the 2002 promotional examination for sergeant was to consist of a written test (weighted at 40% of the total score), an assessment center (32%), a performance review system (8%), and the required education and experience component (20%). The assessment center for sergeants consisted of a "situational exercise using multiple scenarios with oral responses." (Trial Ex. 151 at 3.) The performance review portion was to be based on a candidate's work history. However, the City's collective bargaining agreement with the Boston

Police Patrolmen's Association, the union representing patrol officers (and thus all the candidates for promotion to sergeant), prohibited the Department from using past performance as a factor in evaluating a candidate's qualification for promotion. As a result, Commissioner Davis removed the performance review portion from the exam protocol. The examination as administered consisted of a written test (40 points), an assessment center (40 points), and education and experience (20 points).

Morris & McDaniel was paid more than $1.2 million to develop and administer the 2002 promotional exams for sergeant, lieutenant, and captain. In addition, other expenses associated with the 2002 examination process included transporting, housing, and training a substantial number of police officers from throughout the county who acted as the assessors for the purpose of the assessment center exercise.

Based on the results of the 2002 promotional exam, BPD promoted 69 candidates to sergeant, consisting of 58 whites, 9 African-Americans, and 2 Hispanics. Even with the inclusion of the assessment center in the 2002 exam process, there remained a substantial difference in the promotion rates of minority and non-minority candidates.

At trial, one of the plaintiffs' experts, Dr. Cassi Fields, suggested that the adverse impact of the total scores could have been reduced if, instead of weighting the written exam and the assessment center equally at 40 points, the exam were to be weighted at 20 and the assessment center at 60 points. There is an initial problem with simply altering the weights by fiat; such jiggering with the numbers would be proper only if the revised 20/60 formula could be assessed to be just as valid as the 40/40 formula, and there is no evidence on that point. In any event, one of the defense experts, Dr. Silva, testified that he had tested Dr. Fields' proposition by comparing the test results using the two alternate weighting possibilities and found that the number of

minorities promoted did not change. Using the 40/40 formula, 9 blacks and 2 Hispanics were promoted, while using the 20/60 formula, 8 blacks and 3 Hispanics would have been promoted, but the total minority promotions in either case would have still been 11. Dr. Silva also testified that when he eliminated the oral component entirely and gave the written and E&E components their usual weights of 80% and 20%, 10 minorities would have been promoted, compared to the 11 minorities who were promoted under the original test. In other words, the addition of the assessment center component in the 2002 resulted in one additional minority promotion to sergeant than would have occurred under a test consisting of only the written exam and the E&E component. That is a rather small payoff for the effort and money expended to achieve it.

In summary, Boston made a substantial effort, including the expenditure of well over a million dollars, in devising its 2002 sergeants promotional exam under a delegation from HRD to supplement the written job knowledge test with other elements, including an oral exercise, with the objective of mitigating adverse impact on minority test-takers, but the effort did not succeed.

3. Planning for 2005 and 2008 Exams

When the 2005 and 2008 promotional exams were being contemplated, BPD was operating under budgetary constraints. Former Police Commissioner Kathleen O'Toole testified that she favored the use of an assessment center, but there were no funds available for that purpose. Expense can be a legitimate consideration in evaluating the use of alternatives to a written exam. See Watson, 487 U.S. at 998 (Opinion of O'Connor, J.) ("Factors such as the cost or other burdens of proposed alternative selection devices are relevant in determining whether they would be equally as effective as the challenged practice in serving the employer's legitimate business goals.") The Department adopted various measures in order to deal with budget issues, including offers of early retirement as a means to avoid the necessity of layoffs of younger

officers, and the reduction or elimination of various programs, including its cadet and mounted unit programs. Expending funds to design and administer exams with assessment centers or other additions to the written test under delegation authority as in 2002 was unrealistic under existing budgetary conditions. The same was true in 2008.

Also significant was the failure of the 2002 test, which included an assessment center to supplement the written exam, to have resulted in a material improvement in minority promotions. That exam had been carefully designed and implemented. It was at least doubtful that expending a similar effort would have had greater success in 2005, especially when there may have been a financial incentive to cut corners on the design and execution of such an exam. In brief, the significant cost to design and administer the 2002 exam including an assessment center was not in retrospect supported by gains in the rate of minority promotions. Consequently, for the 2005 and 2008 exams, Boston returned to the prior practice of using the significantly less costly HRD prepared exams, supplemented as usual by Boston-specific questions.

C. Banding

One other potential alternative merits some discussion: "banding." Under the Massachusetts Civil Service regime, once an eligibility list has been generated based on the final scores on the examination, a municipal employer ordinarily must promote candidates in strict rank order in accordance with the list. Thus, a candidate who has scored 89 on the exam will be appointed before a candidate who has scored 88. Critics of this strict procedure argue that for real world work purposes, persons who score one point apart on a written exam are effectively equally qualified and that effective equivalence should be realistically recognized in the promotion process.

Banding would treat scores within a certain number of points of each other as essentially the same grade. For example, if the band width were to be set at three points, then persons scoring 89, 88, and 87 on the exam would be treated as having the same score. Hypothetically, if non-minority candidates scored 89 and 88 and a minority candidate scored 87, under this banding application, the minority candidate would be selectable for promotion, having the same banded score as the others, whereas in a strict rank order regime, the minority candidate would not be selectable. The desired effect would be to increase the number of selectable minorities and thus, hopefully, the number actually selected. All the testifying experts agreed that banding can be a tool to reduce adverse impact without compromising an "employer's legitimate interest in efficient and trustworthy workmanship." Albemarle, 422 U.S. at 425 (internal quotations omitted).

The prospect of potential banding strategies does not help the plaintiffs, however, because the plaintiffs have not shown that the banding technique was realistically available as an alternative to selection in strict rank order for either of the tests in question. For the 2005 test, it does not appear it was contemplated. In contrast, in 2008 HRD actually proposed to change its practice in scoring exams for the statewide sergeants exam to employ a banding strategy, and it appears BPD supported that proposal. However, application of the banding proposal would have substantially altered the statutory "bypass" procedures, and a justice of the Massachusetts Superior Court enjoined the use of banding to score the 2008 police promotional exams on the ground that such a substantial change in HRD's rules and procedures could only be made in a formal rulemaking proceeding under Section 4 of Chapter 31 of the General Laws (encaptioned

"New or amended rules; hearings; publication"), which had not occurred. Pratt v. Dietl, SUVC No. 2009-01254, Memorandum of Decision and Order on Plaintiffs' Motion for a Preliminary Injunction, Apr. 15, 2009. As a result, banding was not actually available as an option to BPD for the 2008 exam.[13]

D. Conclusions Regarding an Actually Available Alternative Selection Method

To prevail, the plaintiffs must show that in 2005 and 2008 Boston had available to it an alternative method of selecting officers for promotion to sergeant that was as valid as the exams actually administered and that would have had a less discriminatory impact on minority candidates. 42 U.S.C. § 2000e-2(k)(1)(A)(ii); Albemarle, 422 U.S. at 425. The statute "requires [the] plaintiffs to demonstrate a viable alternative and give the employer an opportunity to adopt it." Allen v. City of Chicago, 351 F.3d 306, 313 (7th Cir. 2003) (stating that a "vague and indefinite proposal" may not qualify as an "alternative employment practice" under the statute).

What the plaintiffs have been able to demonstrate is that industrial psychologists have in their toolbox various alternatives or supplements to written multiple choice tests that as a general matter tend to result in less adverse impact on minority groups in promotional testing. What they have not been able to show is that there was a particular alternative selection method available for the years in questions about which it could confidently be said that it would have reduced adverse impact on minority candidates for promotion to sergeant in the Boston Police

[13] Moreover, even if banding had been allowed for the 2008 exam, it is impossible to gauge what its effect might have been in increasing the number of minority promotions. Without banding, under the state statute the employer could bypass the two higher scoring candidates and appoint the minority candidate, but the employer would have to justify its decision to the Civil Service Commission. See Massachusetts Ass'n of Minority Law Enforcement Officers, 748 N.E.2d 455. A justification that the choice was race-conscious (to increase the number of minority sergeants, for example) could be subject to challenge under the Massachusetts Civil Service statute as inconsistent with "basic merit principles." Id. at 461-62. It might also be of dubious validity under federal law. See Ricci, 557 U.S. 557.

Department. The 2002 test, which weighted an oral assessment center equally with the written test, failed to lessen adverse impact to any substantial degree or to increase other than marginally the number of minority promotions. The plaintiffs have not offered any evidence that would warrant a conclusion that repeating that experiment in 2005 or 2008 would likely have – as opposed to might possibly have – reduced adverse impact and increased minority promotions. That is not enough to carry their burden on this issue.

I therefore conclude that the plaintiffs have not carried their burden of demonstrating by the evidence that there was an alternative employment practice with equal validity and less adverse impact that was available and that BPD refused to adopt. See 42 U.S.C. § 2000e-2(k)(1)(A)(ii). BPD is entitled to prevail on the plaintiffs' disparate impact claims.

VIII. Conclusions and Order for Judgment

For all the foregoing reasons, I conclude that none of the individual plaintiffs has proved at trial his or her claim for disparate impact discrimination under 42 U.S.C. § 2000e-2(k). Though the reasons vary as to different defendants, all defendants are entitled to have judgment entered in their favor.

It is SO ORDERED.

/s/ George A. O'Toole, Jr.
United States District Judge