UNITED STATES DISTRICT COURT
DISTRICT OF MASSACHUSETTS

BRUCE SMITH, PAUL JOSEPH, JOHN M. JOHNSON, ROBERT TINKER, MARTIN JOSEPH, KIM GADDY, BRIAN KEITH LATSON, LEIGHTON FACEY and MARWAN MOSS, Plaintiffs,	)	CIVIL ACTION NO. 12-10291-WGY
v.	)	
CITY OF BOSTON, Defendant.	)	

YOUNG, D.J. November 16, 2015

FINDINGS OF FACT, RULINGS OF LAW, AND ORDER

I. INTRODUCTION

In this action, ten black police sergeants (the "Plaintiffs") employed by the Boston Police Department (the "Department") brought suit against the City of Boston ("Boston" or the "City") under Title VII of the Civil Rights Act of 1964, alleging that the multiple-choice examinations the Department administered in 2005 and 2008 to select which sergeants to promote to the rank of lieutenant had a racially disparate impact on minority candidates and were insufficiently job-

related to pass muster under Title VII.[1] The Plaintiffs also asserted a pendent claim under Massachusetts General Laws Chapter 151B ("Chapter 151B").[2] The City disputes that the exams had a disparate impact on minority candidates and claims that, even if they did, the exams were sufficiently job-related to survive a Title VII challenge.

This is a profoundly important case, one that evokes the finest of our nation's aspirations to give everyone equal opportunity and a fair shot. In deciding this case, the Court first emphasizes what this case is not about: this is not a case about conscious racial prejudice. Rather, the Plaintiffs' case is rooted in their allegation that the seemingly benign multiple-choice examination promotion process, while facially neutral, was slanted in favor of white candidates.

---

[1] Minority police officers include black and Hispanic officers. Ex. 47, Adverse Impact Evaluation: 2008 and 2005 Exams Promotion Lieutenant, BPD ("Adverse Impact Evaluation") 6. This demographic information is self-reported. 12/15/14 Bench Trial Tr. ("Tr.") 61:6-20, ECF No. 161. Self-reporting makes sense. This Court struggled with the concept of race in Cotter v. City of Boston, 193 F. Supp. 2d 323, 328 (D. Mass. 2002) aff'd in part, rev'd in part, 323 F.3d 160 (1st Cir. 2003).

[2] The legal analysis of disparate impact claims under Massachusetts General Laws Chapter 151B and Title VII is the same. The discussion here will therefore focus on the better developed principles of Title VII jurisprudence, which apply to both claims. See, e.g., White v. Univ. of Mass., 410 Mass. 553, 557 (1991).

The parties engaged in a ten-day bench trial and submitted exhaustive post-trial briefs. The long trial involved substantial and dense discussions of statistical analysis. Consequently, the decision that follows is admittedly complex, but its conclusion is simple: the Department's lieutenant-selection process -- ranking candidates for promotion based on their scores on an exam administered in 2008 ("2008 exam") -- had a racially disparate impact and was not sufficiently job-related to survive Title VII scrutiny. Accordingly, the Court imposes liability on the City.[3]

II. PROCEDURAL HISTORY

The Plaintiffs initiated this case in federal court in February 2012. Compl., ECF No. 1. Judge Tauro, to whom this case was originally assigned, dismissed without prejudice the claims of two of the Plaintiffs (John Johnson and Robert Tinker) for their failure to exhaust administrative remedies. Mem., ECF No. 28. Once this case was transferred to this Session on December 26, 2013, Mem., ECF No. 56, this Court denied the Plaintiffs' motion to reconsider the dismissal, Elec. Order, ECF No. 67, and subsequently denied without prejudice the Plaintiffs' motion to certify a class, Elec. Clerk's Notes, ECF

[3] The Court does so without deciding whether the Plaintiffs have satisfactorily proven a less discriminatory alternative. See infra note 43 (explaining then rejecting the City's argument that the Court must decide the issue as a prerequisite to imposing liability on the City).

No. 70. Two years of discovery ensued, followed by the virtually inevitable cross-motions for summary judgment. Def.'s Mot. Summ. J., ECF No. 89; Pls.' Mot. Summ. J., ECF No. 94. This Court denied summary judgment on all claims due to genuine disputes of material fact. Elec. Clerk's Notes, ECF No. 120. In December 2014, the Court ruled that the remaining eight Plaintiffs had no viable disparate impact liability claim arising from their taking the 2005 lieutenant promotional exam (the "2005 exam") due to their failure to exhaust administrative remedies. Elec. Order, ECF No. 150. Although no longer formally the subject of this litigation, the Court did consider evidence regarding the 2005 exam for background and context in evaluating the 2008 exam.

At the pre-trial conference, the Court bifurcated the case into separate liability and damages phases. Elec. Clerk's Notes, ECF No. 98. The liability phase was tried before the Court between December 15, 2014 and January 7, 2015. See 12/15/14 Tr. 3:3-4, ECF No. 161; 01/07/15 Tr. 3:9-11, ECF No. 160. The following witnesses testified for the Plaintiffs: Dr. Joel Wiesen, PhD., industrial organizational psychologist (expert witness), 12/15/14 Tr. 3:6-12, Department Sergeant and Plaintiff Bruce Smith (fact witness), 12/17/14 Tr. 3:11-13, ECF No. 163, former Department Commissioner Edward Davis (fact witness), 01/05/15 Tr. 3:9-11, ECF No. 158, and Leatta M. Hough,

PhD, industrial organizational psychologist (expert witness), 01/06/15 Tr. 3:5-7, ECF No. 159. The following witnesses testified for the City: Dr. Jacinto Silva, PhD, industrial organizational psychologist (expert witness), 12/17/14 Tr. 3:15-17, Dr. Michael Campion, PhD, industrial organizational psychologist (expert witness), 12/19/14 Tr. 3:5-7, ECF No. 166, Department Chief of the Bureau of Administration and Finance Edward P. Callahan (fact witness), 01/06/15 Tr. 3:9-11, and Department Commissioner William E. Evans (fact witness), 01/07/15 Tr. 3:5-7.

III. LEGAL CONTEXT

A. Title VII

It is the goal of Title VII "that the workplace be an environment free of discrimination, where race is not a barrier to opportunity." Ricci v. DeStefano, 557 U.S. 557, 580 (2009). The statute is designed to "'promote hiring on the basis of job qualifications, rather than on the basis of race or color.'" Id. at 582 (quoting Griggs v. Duke Power Co., 401 U.S. 424, 434 (1971)).

Title VII, codified at 42 U.S.C. § 2000e, provides two theories of liability for discrimination in the employment context: disparate treatment and disparate impact. Ricci, 557 U.S. at 577-78. A disparate treatment claim accuses an employer of intentionally basing employment decisions on an improper

classification, such as race. See id. at 577. A disparate impact claim, by contrast, challenges an employment decision that is facially neutral, but which falls more harshly on those in a protected class. See id. at 577-78.

This is a disparate impact case. Second Am. Compl., Compensatory, Injunctive & Declaratory Relief Requested (the "Complaint") ¶ 1, ECF No. 14. Section 2000e-2(k)(1)(A) outlines the burden of proof in a disparate impact case:

> An unlawful employment practice based on disparate impact is established under this subchapter only if –
>
> (i) a complaining party demonstrates that a respondent uses a particular employment practice that causes a disparate impact on the basis of race, color, religion, sex, or national origin and the respondent fails to demonstrate that the challenged practice is job related for the position in question and consistent with business necessity; or
>
> (ii) the complaining party makes the demonstration described in subparagraph (C) with respect to an alternative employment practice and the respondent refuses to adopt such alternative employment practice.

42 U.S.C. § 2000e-2(k)(1)(A).

Under First Circuit case law, the plaintiff bears the burden of establishing a prima facie case of discrimination which consists of identification of an employment practice (in this case, the 2008 exam and promotions flowing therefrom),[4]

[4] Identifying an employment practice can be a tricky proposition in the context of Title VII. As is typical in cases

disparate impact, and causation. Bradley v. City of Lynn, 443 F. Supp. 2d 145, 156 (D. Mass. 2006) (Saris, J.) (quoting EEOC v. Steamship Clerks Union, Local 1066, 48 F.3d 594, 601-02 (1st Cir. 1995)).

If the Plaintiff meets this burden, the employer may either debunk the Plaintiff's prima facie case, or alternatively, may demonstrate that the challenged practice is "job-related and consistent with business necessity." Bradley, 443 F. Supp. 2d at 157; see also Ricci, 557 U.S. at 578. If the employer demonstrates the latter, the ball bounces back into the plaintiff's court to demonstrate that "some other practice, without a similarly undesirable side effect, was available and would have served the defendant's legitimate interest equally well." Bradley, 443 F. Supp. 2d at 157.

The law of disparate impact has become a powerful tool for ensuring equal opportunity. It is both balanced and nuanced. Each step in its three-part doctrine serves a valuable function. Consider the following.

---

involving a battle of statistics, the parties argue that the statistic more favorable to them is the appropriate one. The Plaintiffs here argue that the Court may and should consider promotion rates, pass-fail rates, average scores, and delays in promotion stemming from the 2008 exam. The City argues that promotion rates are the only pertinent statistic. As is explained below, the Court sides with the Plaintiffs on this disagreement. See infra Part V-A-1.

All testing, hiring, and promotion schemes are necessarily discriminatory. These programs exist because there are more applicants than there are jobs. Under the first prong, the Plaintiffs must make a significant showing of actual disparate impact upon an identified protected minority. This is as it should be: no one wants federal courts acting as super personnel agencies.

If the plaintiff can, however, make this showing, then under the second prong, the employer gets a chance to demonstrate that the test in question is both job-related and consistent with business necessity. Again, this step is sensible: courts ought not defer excessively to employers, but neither should they ignore the realities of particular jobs. The current debate over the exhaustive testing to determine the capabilities of women to engage in military ground combat comes immediately to mind as exemplifying the difficult issues encountered in prong two.

Even if the employer succeeds, however, the case is not over. Under the third prong, the plaintiff gets one more shot. If the plaintiff can demonstrate the availability of a testing program equally determinative of job performance, yet resulting in less disparate impact, the Court should fashion a remedy to secure the greatest degree of equal opportunity. In other words, to produce more equality of opportunity, Title VII

empowers courts to impose on employers an equally effective means of evaluating applicants.

B. The Commonwealth's Statutory and Administrative Framework

Under Massachusetts law, police sergeants seeking to be promoted to lieutenant are subject to the state civil service statutory promotion regime. See Mass. Gen. Laws ch. 31, § 51. The purpose of the examination regime is 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, 304 (1997). Under this regime, to become a lieutenant, Boston police sergeants must first pass a competitive civil service examination. Mass. Gen. Laws ch. 31, § 59.

The Commonwealth of Massachusetts Personnel Administrator of the Human Resources Division ("HRD"), Mass. Gen. Laws ch. 31, § 1, is responsible for "conduct[ing] examinations for purposes of establishing eligible lists" for promotion. Id. § 5(e). HRD is obligated by statute to "fairly test the knowledge, skills and abilities which can be practically and reliably measured and which are actually required" to perform the job. Id. § 16. To achieve this end, HRD develops the examination, id. §§ 5(e), 16, posts notices of the exams, id. §§ 18-19, and determines the

passing requirements, id. § 22. Promotional examinations within the Department are typically administered every two or three years. 12/15/14 Tr. 64:4-10, ECF No. 161. Unlike other jurisdictions, Massachusetts requires candidates for each supervisory rank (sergeant, lieutenant, and captain) to take a test at each promotional level, even if there is substantial overlap in the questions that appear on the tests for each rank. Id. at 129:19-130:5.

The Department has the option of either using the exams developed by HRD, or seeking a delegation agreement with HRD by which HRD agrees to oversee the Department creating its own exam. Mass. Gen. Laws ch. 31, §§ 5(l), 59, 65; 01/06/15 Tr. at 91-93. Under this delegation regime, municipalities must still comply with civil service law and regulations, but may decide themselves how to satisfy these requirements. Lopez v. Massachusetts, 588 F.3d 69, 76 (1st Cir. 2009). When the City enters into such a delegation agreement, it must bear the costs of developing and administering the tests. 12/17/14 Tr. 49:21-24. The City does not incur these costs when it uses an HRD test. 01/06/15 Tr. 112:16-21.

In a competitive examination, "an applicant shall be given credit for employment or experience in the position for which the examination is held." Mass. Gen. Laws ch. 31, § 22. Such credit is known as an education and experience score ("E&E"),

and is calculated using biographical information provided by the applicants. Id.; Ex. 85, Affidavit Edward P. Callahan ("Callahan Aff."), ECF No. 177; Ex. 3, Education Experience Sheet Instructions ("E&E Instructions") 1, ECF No. 177-3. Pursuant to Massachusetts statute, veterans and long-service employees receive preference points. Mass. Gen. Laws ch. 31, §§ 26, 59.

After combining the exam scores with the E&E scores, HRD issues an eligibility list for specific positions ranked in order of an applicant's total score. Id. §§ 25, 27.[5] An eligibility list remains in effect until replaced by a new eligibility list from a subsequent exam. Id. § 25; Callanan v. Pers. Adm'r for the Commonwealth, 400 Mass. 597, 601-02 (1987). The City has promoted police officers based on promotional examination results in strict rank order since at least 1977. 12/15/14 Tr. 63-64.

To hire for a vacancy, the Department submits a request to HRD, which certifies from the larger eligibility list a smaller list of names of persons for consideration in rank order. Mass. Gen. Laws ch. 31, § 6. Under HRD's Personnel Administration

[5] Section 26 seems to mandate that other categories of qualified applicants (meaning those who "pass[ed the relevant] examinations"), including disabled veterans, veterans, and widows or widowed mothers of veterans who were killed in action, be placed above even the highest scorers on the test. Mass. Gen. Laws ch. 31, § 26.

Rules, the number of candidates appearing on the smaller list is determined by the formula 2n+1, with n representing the number of vacancies. 12/15/14 Tr. 63:5-16. For example, if there are two job vacancies corresponding to one applicable list, the list would be comprised of the candidates with the five highest scores (2x2+1=5).

The Department must then make selections from that list based on strict rank order, based on the candidates' performance on the promotional exam. 01/06/15 Tr. 102:12-18. If the candidates have tied scores, the Commissioner may consider factors such as past work history and diversity. 01/07/15 Tr. 9:25-11:19. The statutory framework allows a municipal employer to "bypass" a candidate on the list -- that is, to step out of strict rank order -- but the employer must have a defensible reason for the bypass, such as a history of disciplinary infractions. Mass. Gen. Laws ch. 31, § 27; City of Cambridge v. Civil Service Comm'n, 43 Mass. App. Ct. 300, 305 (1997). Former Commissioner Davis testified that as a practical matter, bypassing is difficult. 01/05/15 Tr. 100-101.

Dissatisfied candidates may challenge the results of the examination with the HRD by claiming that the examination was not 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." Mass. Gen. Laws ch. 31, § 22. The

Massachusetts Civil Service Commission oversees the administrative appeals process for candidates to air their grievances with the hiring process and the initial review by the HRD. Id. § 24. Judicial review is available after the candidate has exhausted his or her administrative remedies. Id. § 44.[6]

IV. FINDINGS OF FACT

This is not the first case in which Department employees or potential employees have challenged the Department's hiring or promotional procedures. In fact, a case raising similar issues to this one was brought before Judge O'Toole in 2007. Judge O'Toole's findings of fact and conclusions of law issued in September 2014. Findings Fact, Conclusions Law, Order J., Lopez v. City of Lawrence ("Lopez"), ECF No. 347 (D. Mass. Sept. 5, 2014). An appeal is pending. Lopez, appeal docketed, No. 14-01952 (1st Cir. Sept. 17, 2014). As there is some factual overlap, on the first trial day in this case, without objection, this Court admitted in evidence all the trial testimony and exhibits from Lopez. 12/15/14 Tr. 32:4-16.[7]

[6] The Plaintiffs remaining in this case have done so.

[7] In Lopez, various minority police officers across the Commonwealth who were vying for a promotion to become sergeants challenged the promotional exams under state and federal law. This Court's conclusions differ from the Lopez decision in one crucial respect. See infra note 42 and accompanying text.

More broadly, over the past few decades, candidates who are members of a racial minority and candidates who are not have challenged the hiring and promotional procedures employed by the Department as unlawfully discriminatory. The extensive litigation history is well documented elsewhere, see Lopez at 12-14, and the Court will not repeat it here. The Court mentions it only for the purpose of noting that, similarly to police departments in other jurisdictions, this pendulum of litigation has pressured the City constantly to re-tool its examination procedures in an effort to appease job applicants and courts alike. See Barnhill v. City of Chicago, Police Dep't, 142 F. Supp. 2d 948, 949 (N.D. Ill. 2001).

A. The Role of a Boston Police Department Lieutenant

Boston Police Department lieutenants are second-line supervisors, meaning that they supervise sergeants, who themselves supervise police officers out in the field. 01/05/15 Tr. 102:1-10. Lieutenants are also in charge of station houses and are responsible for the proper arrest of suspects and for the safety of prisoners. Id. at 128:3-9, 131:4-8. The job involves a significant amount of desk work, 12/17/14 Tr. 104:9-12, as well as work outside of the station, including talking with citizens at community meetings, id. at 105:14-25, and taking control of scenes of major incidents, 01/05/15 Tr. 126:23-127:12. The job requires good management skills,

including the ability to motivate employees, and to communicate information between ranks. Id. at 102:11-23.

The official Department job description for lieutenant has not changed since 1979, and current Boston Police Department Commissioner William Evans testified that it remains accurate today. Ex. 23, Boston Police Department Rules Procedures, Rule 105; 01/07/15 Tr. 16:15-17:9. Just prior to 2006, however, Boston began to shift its policing philosophy towards that of community policing, where police officers engage more directly with the community they serve. 01/05/15 Tr. 109:16-110:15. The skills required for a Boston Police Department lieutenant evolved with this shift, differing from what was needed in the early 1990s when the prevailing policing philosophy involved responding to, rather than preventing, emergencies. Id. at 110:9-12.

B. Job Analyses and Validation Studies Pre-Dating 2005

The first step in developing a valid civil service promotional exam is to create a job analysis, which identifies "important work behavior(s) required for successful performance and their relative importance." 29 C.F.R. § 1607.4(C)(2) ("Uniform Guidelines").[8]

[8] Chapter 29 of the Code of Federal Regulations, section 1607, was published under the name of Uniform Guidelines on Employee Selection Procedures in 1978 by several government agencies to interpret how selection and testing and assessment

The development of the 2005 and 2008 exams began in 1991 with the creation of a job analysis and validity report, which HRD incorporated into the 2008 exam. 12/19/14 Tr. 14:5-15. HRD also incorporated into the 2008 exam a job analysis from 2000, which was in part based on the 1991 job analysis.[9] Id. at 14:5-15:25, 32:13-16. The documents are somewhat dense, and the Court will discuss each in turn.

1. The 1991 Job Analysis

In 1991, the Massachusetts Department of Personnel Administration ("DPA"), the predecessor to HRD, prepared a state-wide validation report for the ranks of sergeant, lieutenant, and captain.[10] Ex. 71, Validation Report 1991 Police Promotional Selection Procedure ("1991 Validation Report") at 00247. DPA relied on the Uniform Guidelines, as well as the Society for Industrial and Organizational Psychology Principles for the Validation and Use of Personnel Selection Procedures ("SIOP Principles"). Id. at 00247, 00250.

---

should be conducted in accordance with the Civil Rights Act of 1964. 12/19/14 Tr. 23:6-18.

[9] The parties disagree as to whether the exams were based on both job analyses, or on the 2000 job analysis alone. The 2000 report explicitly states that it reviewed "past job analyses" to prepare the 2000 report. Ex. 39, Job Analysis Report Police Lieutenant City Boston 9. The Court regards both job analyses as relevant to this case.

[10] In the Department, captains supervise lieutenants, and lieutenants supervise sergeants. Ex. 46, Boston Police Organizational Structure 2.

DPA began the job analysis by gathering information about the positions of police sergeant, lieutenant, and captain. It did so by surveying other jurisdictions and reviewing various documents including job analysis studies, articles, and the Uniform Guidelines. 1991 Validation Report at 00253-54. Based on this research, DPA created a list of 136 potentially critical tasks that sergeants, lieutenants, and captains perform. Id. at 00256. DPA sent surveys to municipal police departments across the Commonwealth asking incumbents of the positions to identify which tasks were critical to successful job performance and then rate them accordingly. Id. Police officers were asked to provide information regarding how often they performed the tasks, and to identify the fifteen tasks most critical to their jobs. 1991 Validation Report, App. H at 3638.

DPA then used these task ratings to create a list of 187 Knowledge, Skills, and Abilities ("KSAs") necessary to carry out the critical tasks. 1991 Validation Report at 00257-58. In a survey administered to subject matter experts ("SMEs") -- superior officers serving in Massachusetts police departments -- DPA asked the SMEs to evaluate two things: the importance of the KSAs, and whether the candidate needed the KSA at the time of appointment to the position or could acquire the KSA on the job. Id. Only KSAs that were determined to be important and

necessary upon starting the job were considered for inclusion in the test for applicants. Id. at 00258.

In addition to these various surveys, DPA developed additional KSAs by holding "critical incident" technique discussions with the SMEs. Id. at 00252. These discussions involved SMEs providing narrative explanations or anecdotes about the tested positions. Id. The purpose was to allow DPA to design situational examination questions evaluating supervisory abilities. Id.

The next step was to link the KSAs with the critical tasks. Nine SMEs were tasked with this assignment. Id. at 00258. The SMEs identified, by group consensus, which KSAs had a direct relationship to identified clusters of critical tasks. Id. at 00259.[11] Of the 187 KSAs, 60 were ultimately incorporated in the written test. 1991 Validation Report, App. EE.

The 1991 report acknowledged that some of the skills identified as important by SMEs could not be evaluated by a written test, such as the "ability to establish rapport with persons from different ethnic, cultural, and/or economic backgrounds." Id. at 00265. The 1991 report also noted that "assessment of the performance of these skills and abilities would require the use of selection devices outside the scope of

[11] DPA also conducted structured discussions with SMEs to discuss the E&E component. Id. at 00260.

the written, multiple choice format." Id. at 00265. Nevertheless, DPA decided to proceed exclusively with a written examination, offering eighty questions common to all three positions; an additional twenty questions for lieutenants and captains testing knowledge of police supervision, administration, and management; and another twenty-five questions to test captains on their knowledge of police administration. Id. at 00266-67. For the 1991 exam, the written portion accounted for 80% of an applicant's final score, and the E&E portion for 20%. Id. at 00263.

2. 2000 Job Analysis and the Corresponding 2002 Exam

The 2000 job analysis (the "2000 report") was prepared at the request of the City of Boston by Morris & McDaniel, Inc., a consulting firm that specializes in the development of promotional systems. Ex. 39, Job Analysis Report Police Lieutenant City Boston ("2000 Job Analysis Report"). The 2000 report at issue in this case concerned Boston Police Department lieutenants only. Id. Morris & McDaniel came up with a list of 302 possibly relevant tasks that Boston police lieutenants perform, as well as KSAs necessary to carry out those tasks. Id. at 65; id., App. A, Task Inventory Police Lieutenant Boston Police Department.

Morris & McDaniel then had twelve SMEs, consisting of Department employees holding the rank of Lieutenant or higher,

rate the tasks for frequency, importance, necessity of performing the task upon starting the job, and how correlative successful performance of the task was to successful job performance. 2000 Job Analysis Report at 10-14. If ten of the SMEs rated a task as "very important" or "important," necessary upon entry to the job, and agreed that performance of that task clearly separated the best workers or better workers from inferior workers, then it satisfied Morris & McDaniel's test criteria. Id. at 14. Of the initial 302 tasks, 281 fulfilled the criteria. Id. Morris & McDaniel also asked the SMEs to determine which of the following dimensions were required for each task: oral communication, interpersonal skills, problem identification and analysis, judgment, and planning and organizing. Id. at 15. Morris & McDaniel then composed a list of 149 KSAs potentially necessary to perform the 281 tasks. See id. at 48-49. Next, the SMEs were asked whether the KSAs related to the job of police lieutanant, when the KSA was learned (before or after assignment to the job), how long it took to learn the KSA, how the KSA differentiated performance, and whether the KSA was required to perform the job effectively. Id. For a KSA to be important enough to be tested, nine of the twelve SMEs must have rated the KSA as related to the job, learned before assignment to the job, requiring more training than a brief orientation period, capable of distinguishing

performance to a high or moderate degree, and required or desirable to perform the job effectively. Id. at 49. Of the 149 KSAs rated by the SMEs, 145 were deemed sufficiently important to be tested. Id.

Based on its 2000 job analysis, Morris & McDaniel created a promotional test for the Department, to be administered in 2002. Ex. 80, Draft Police Lieutenant Written Examination Validity Report ("Validity Report") 1.[12] The 2002 exam consisted of three components: a written exam, an assessment center, and E&E. Id. at 2.[13] Of a possible 100 points, Morris & McDaniel, after consulting with SMEs, assigned 30% to the written examination, 50% to the assessment center, and 20% to E&E. Id.; Ex. 81, Police Lieutenant Assessment Center Validity Report ("Assessment Center Report") 17; 12/22/14 Tr. 46-47, ECF No. 167; 01/06/15 Tr. 110:7-13.

[12] A draft of the validity report for the written examination was admitted in evidence because counsel for the Plaintiffs indicated that they could not locate a final copy. 12/22/14 Tr. 39:14-40:8. Neither party objected to its admission or argued that the final report contained conclusions different from the draft.

[13] As part of the 2002 exam, the Department attempted to introduce a performance review. In response, the police officers' union brought a complaint to the Civil Service Commission. 01/06/15 Tr. 88:17-21. The Department ultimately excluded the performance review system from the 2002 exam. Id. at 88:23-25.

Morris & McDaniel composed the written test questions, which SMEs reviewed for accuracy and clarity. Validity Report 11-12. Based on the responses, Morris & McDaniel determined that the questions were internally consistent and reliable. Id. at 13.

The assessment center was designed to test oral communication skills, interpersonal skills, ability to quickly identify a problem and analyze it, ability to make sound decisions promptly, and ability to break work down into subtasks and prioritize them.[14] Assessment Center Report 5-6. The assessment center consisted of an in-basket exercise (a simulated written exercise), and a situational exercise in which candidates were videotaped offering verbal responses to hypothetical scenarios that a lieutenant may encounter. Id. at 7-8. The assessment center exercises were evaluated by outside assessors. 01/06/15 Tr. 110:14-23.

After the 2002 exam had been administered, Morris & McDaniel prepared validity reports of the written examination and the assessment center. See Validity Report; Assessment Center Report. The validating report was written to comply with the Uniform Guidelines. Validity Report 1. Morris & McDaniel

[14] Testimony in the Lopez case indicated that the use of the assessment center was in part the result of an effort by the Department to increase diversity among its promotional ranks. Lopez 07/27/10 Trial Tr. 100.

concluded that both portions of the examination were valid. Id. at 18.

This process cost $1,300,000, which included consulting fees and travel costs for outside assessors. 01/06/15 Tr. 112-113. HRD certified a list for promotion, from which one black sergeant was selected for promotion. Ex. 38; 01/06/15 Tr. 112:2-4.

C. Development and Administration of the 2008 Exam

Based in part on financial constraints and a perceived lack of improvement in diversity resulting from the 2002 exam, the Department elected to use HRD exams (conventional, written exams) in 2005 and 2008.[15] Lopez 07/28/10 Tr. 17, 30; 01/06/15 Tr. 93:9-12, 118-19.[16] HRD consulted with the outside firm EB Jacobs for the 2008 exam. 12/16/15 Tr. 55:17-56:3.

HRD did not create a comprehensive job analysis for the 2008 exam, but instead conducted "mini job analyses" for all three ranks, which were more or less updates or overlays to the 2000 report. HRD asked SMEs to rate tasks and KSAs. Ex. 55;

[15] The development and administration of the 2005 and 2008 exams were substantially similar. In an effort to avoid redundancy, the Court focuses on the 2008 exam because it forms the crux of this dispute.

[16] Then-police commissioner Kathleen O'Toole wanted to include an assessment center component in the 2005 exam, a request that was ultimately rejected because of funding concerns. 01/06/15 Tr. 113-114.

Ex. 56; Tr. 01/07/15 at 20-36. The KSAs were pulled from the 2000 job report. 12/19/14 Tr. 38:9-19.

Commissioner Evans, who was a fact witness in this case, served as an SME for the 2008 exam. 01/06/15 Tr. 126:6-126:10.[17] He testified that the SME review of the KSAs and tasks was "meticulous," 01/07/15 Tr. 30:25-31:4, and that the purpose of the 2008 exam was not to test memorization of facts, but to test situational judgment, interpersonal relations, communication ability, and knowledge of rules and regulations. 01/07/15 Tr. 35-36.

Based on the mini job analyses, the Department's consultant, EB Jacobs, created a test outline for the 2008 exam. Ex. 54; Ex. 60.[18] The Civil Service then compiled a list of 100 questions for the exam. 01/07/15 Tr. 31:5-10. The SMEs reviewed the test questions for suitability for the different ranks, difficulty, and readability, and then indicated whether they recommended the question for the exam. Ex. 60; 01/07/15 Tr. 31:11-17.

HRD announced the 2008 exam and provided a corresponding reading list to members of the Department. Exs. 1, 5, 6, 8, 17.

[17] The other SMEs for the 2008 exam were Captain Purvis Ryan, Captain Mark Hayes, and Captain Genevieve King. 01/06/15 Tr. 125:6-126:10.

[18] Trial exhibits without clearly identifiable titles are referred to simply by their numbering.

The Department provided "substantial amount[s] of tutorial information" for all candidates taking the promotional exams, including taped lectures and practice questions. 01/06/15 Tr. 128. The questions that appear on the written portion of the exam are taken directly from the reading list. 12/15/14 Tr. 51:19-52:12.[19]

The 2008 exam consisted of two elements: a written, closed-book exam consisting of 100 multiple-choice questions, and an E&E rating. 12/15/14 Tr. 49:7-14. Out of 100 possible points on the written examination, a candidate needed to score 70 to pass. Id. at 62:21-23. The E&E Score is calculated only for candidates who passed the written exam. Id. at 62:24-63:2. The written portion accounted for 80% of the final score; the E&E component for 20%. Id. at 50:9-50:15.

D. Development and Administration of the 2014 Exam

The Department typically offers promotional exams every two or three years. Id. at 64:4-7. The 2008 exam did not follow this trend: Commissioner Davis requested that the promotional certifications be extended due to the Lopez litigation. Id. at 64:11-18; Ex. 59; Ex. 61. Thus, it was not until 2014 that the Department developed and administered a new promotional exam;

[19] The topics covered by the reading list for the 2008 exam included police supervision, management, administration, and other aspects of Massachusetts law. 12/17/14 Tr. 134-36.

the list from the 2008 exam was still in effect at the time of this trial. 01/06/14 Tr. 143:25-144:3.

The Department, in part due to an improved economic climate, and in part due to a desire to increase diversity, elected to go beyond a written examination and E&E for the 2014 promotional process. 01/05/2015 Tr. 139-140:8. The Department again retained the firm of EB Jacobs, this time to design and administer the 2014 exam. Callahan Aff. ¶ 1. At the firm's recommendation, the Department in 2013 approved the development of a "very comprehensive" job analysis in anticipation of the 2014 exam ("2013 job analysis"). 01/06/15 Tr. 109:11-19. This was the first full job analysis performed since 2000. Id. at 131:5-19. Based on the 2013 job analysis, EB Jacobs recommended the use of an assessment center in addition to a written examination. Id. at 133:12-23. After securing funding from Boston for an assessment center, the Department secured a delegation from HRD to develop its own promotional exam. Id. at 134:5-17.

The Department posted an announcement of the exam, and indicated that the exam would be weighted as follows: technical knowledge written exam (36%); in-basket test (where candidates provide written essay-style responses to various job situations) (20%); oral board test (where candidates provide oral responses to hypothetical incidents and personnel issues) (24%); and E&E

(20%). Callahan Aff. ¶¶ 4, 7; id., Ex. 2, Lieutenant Promotion Examination Candidate Preparation Guide In-Basket Oral Board Tests 3, 5, ECF No. 177-2.

Unlike the 2005 and 2008 exams, there was no cut-off score for the written portion of the 2014 exam. 12/17/14 Tr. 111:25-112:2. Much else was the same, however. Outside assessors evaluated each candidate's performance in the assessment centers, id. at 114:22-115:4, the E&E component was based on self-reporting of candidates' education, training and work experience, E&E Instructions 1, and veterans received an additional two points, Callahan Aff. ¶ 9. The full development of the promotional exam process (testing for promotion of sergeant, lieutenant, and captain) cost over $1,600,000. Id. ¶ 12.

E. Results of the 2005 and 2008 exams

While all of this history is informative and helpful to gain an understanding of the Department's promotional process, it is the results of the 2005 and 2008 exams that form the crux of this dispute. By and large, the parties agree on all of the numbers in this section. The crux of the dispute is which of these numbers are important, and which methodology is the most appropriate for analyzing these numbers.

One hundred and twenty seven candidates reported for the 2005 lieutenant promotional exam: 104 were white, 22 black, and

1 Hispanic.[20] Ex. 47, Adverse Impact Evaluation: 2008 and 2005 Exams Promotion Lieutenant, BPD ("Wiesen Report") 8; Ex. 72, Report Jacinto M. Silva ("Silva Report") 2. The passing rate of the written exam for minorities was 50%, and for whites, 88%. Wiesen Report 12. The mean score for minorities on the 2005 exam was 69.9, and for whites, 78.7. Id. at 16. Of the 127 sergeants who sat for the exam, 27 were promoted: 25 were white (out of 104 white applicants), one was black (out of 22), and one was Hispanic (who was the only Hispanic candidate). Id. at 8; Silva Report 2.

Ninety-one sergeants sat for the 2008 promotional exam: 65 were white, 25 were black, and one was Hispanic. Wiesen Report 7; Silva Report 7. Of the 91 candidates who took the exam, the passing rate for minorities was 69%, and for whites was 94%. Wiesen Report 11. The mean score for minorities was 76.6, and for whites was 83.2. Id. at 14. Of the 91 candidates, 33 were promoted: 28 were white (out of 65), and 5 were black (out of 25). Silva Report 7.

[20] The expert reports on disparate analysis (Wiesen Report, Silva Report) disagree as to whether the number of people who took the 2005 promotional exam was 126 or 127. The parties agree that the Court should adopt 127, the number more favorable to the Plaintiffs, and also agree that the difference will have only a minor impact. Proposed Findings Fact City Boston ("Pre-Trial City Proposed Findings") ¶ 25, ECF No. 141; Silva Report 2.

After reviewing the final scores from the 2008 exam, the Department's consultant, EB Jacobs, recommended "banding" the results in nine-point increments (meaning that scores within nine-point ranges would be deemed equivalent). Lopez Exs. 70, 71.

The background and raw numbers for this case are straight-forward and self-explanatory. Their statistical analysis and corresponding legal conclusions, less so.

V. CONCLUSIONS OF LAW

A. Disparate Impact (Prong 1)

The use of the 2008 exam is the employment practice subject to challenge under Title VII. The parties agree that minority test-takers passed the 2008 exam and were promoted to lieutenant at a lower rate when compared to white candidates. But a lower rate of passage or promotion is not, by itself, sufficient to establish a prima facie case of discrimination. Statistical disparities must be significant enough to "raise an inference of causation." Bradley, 443 F. Supp. 2d at 157 (citation omitted). The Supreme Court has stated:

> [T]he plaintiff must offer statistical evidence of a kind and degree sufficient to show that the practice in question has caused the exclusion of applicants for jobs or promotions because of their membership in a protected group. Our formulations, which have never been framed in terms of any rigid mathematical formula, have consistently stressed that statistical disparities must be sufficiently substantial that they raise such an inference of causation.

Watson v. Fort Worth Bank & Trust, 487 U.S. 977, 994-95 (1988) (plurality opinion); see also Texas Dep't of Hous. & Cmty. Affairs v. Inclusive Communities Project, Inc., 135 S. Ct. 2507, 2523 (2015) (noting that "[a] robust causality requirement . . . protects defendants from being held liable for racial disparities they did not create"). In other words, the Plaintiffs must show that any disparity between races is not the result of mere chance. See Jones v. City of Boston, 752 F.3d 38, 43 (1st Cir. 2014).

There is no "single test" to demonstrate disparate impact. Langlois v. Abington Hous. Auth., 207 F.3d 43, 50 (1st Cir. 2000). Plaintiffs in Title VII disparate impact cases often demonstrate causation by presenting evidence that the disparity in outcomes between white and minority candidates is "statistically significant," meaning that statistical analysis reflects that the odds of the disparity occurring by mere coincidence are less than 5%. Jones, 752 F.3d at 43-44. This is demonstrated when a statistician determines that the "p-value," which stands for "probability," is less than .05, meaning that the probability of the result occurring by chance is less than 5%. Jones, 752 F.3d at 46-47; 12/15/14 Tr. 74:19-75:7; Wiesen Report 5.

Another way to express this same mathematical calculation is to utilize the statistical measure of "standard deviation,"

which measures how dispersed a set of data is (the more dispersed the data, the higher the standard deviation). See 12/15/14 Tr. 75:19-25. One "can calculate the standard deviation . . . for any [p-value]." Id. at 77:1-9. For a so-called one-tailed test,[21] the relevant standard deviation is 1.645. 12/15/14 Tr. 77:1-9. For a two-tailed test, it is 1.96. Id. In other words, if one utilized a two-tailed test and found that the mean test scores of minority candidates were located 1.96 standard deviations away from the overall mean, there would be only a 5% probability that such difference was due to chance. (And if their mean score was more than 1.96 standard deviations from the mean, the probability that it was due to chance would be even lower.)

Parties alleging disparate impact also sometimes rely on what is known as the "four-fifths rule," articulated in the Uniform Guidelines, a non-binding set of guidelines authored by the Equal Employment Opportunity Commission to help employers comply with Title VII. The thrust of the "four-fifths rule" is that a selection rate for any racial group that is less than four-fifths (or 80%) of the rate of the group with the highest rate is evidence of adverse impact. 29 C.F.R. § 1607.4(D). The

[21] The Court discusses below at some length the distinction between one-tailed tests and two-tailed tests. See infra Part V-A-3.

four-fifths rule is "widely used" among organizational psychologists. 12/15/14 Tr. 72:25-73:6. The First Circuit has acknowledged, however, that it is not decisive. Jones, 752 F.3d at 51.

Reports and testimony from experts play a crucial role in evaluating a disparate impact claim. The Plaintiffs' expert witness regarding disparate impact was Dr. Joel Peter Wiesen, an industrial organizational psychologist. 12/15/14 Tr. 34:3. Dr. Wiesen holds his PhD in psychology. Id. at 34:11. He worked for HRD from 1977-1992, during which time he was the chief of test development and validation. Id. at 36:13-25. Since leaving HRD, Dr. Wiesen has been consulting in the area of test development, and has served as an expert witness. Id. at 38:10-16.[22] The City rebutted Dr. Wiesen's testimony with that of Dr. Jacinto M. Silva, who holds a PhD in industrial organizational psychology. 12/17/14 Tr. 141:19-142:11; Silva Report 1. Dr. Silva is currently a senior managing consultant at EB Jacobs, which is the firm that developed the 2014 lieutenants' exam for the Department and consulted on the 2008 exam. Id. at 143:25-144:11.

Drs. Wiesen and Silva agreed on many issues: the raw data underlying each other's analysis (although there were some minor

[22] The Plaintiffs also offered Dr. Wiesen's testimony on the issue of test validation. 12/15/14 Tr. at 42:17-22.

discrepancies based on the timing of their reports); that each other's mathematical calculations were correct; and that the Fisher Exact Test was the appropriate test for this analysis. Def. City Boston's Post-Trial Proposed Findings Fact & Conclusions Law ("Post-Trial City's Proposed Findings") 38 n.21, ECF No. 190. The experts disagreed as to which statistics were relevant (among promotion rates, mean scores, pass-fail rates, or delay in promotion); whether the results from the 2005 and 2008 tests should be aggregated; and whether a one-tailed or a two-tailed Fisher Exact Test was the appropriate methodology. The Court will address these issues in turn.

1. The Relevant Data Points

Before beginning any analysis, statistical or otherwise, the Court must determine the proper scope of its inquiry. The Plaintiffs argue the Court should cast a wide net, looking to several statistics regarding the 2008 exam. The City, however, suggests that one number, promotion rates, provides all the necessary information. As explained more fully below, the Court largely agrees with the Plaintiffs.

Dr. Wiesen examined various aspects of the lieutenant promotional procedure employed by the City in an effort to determine whether there was disparate impact. Specifically, he compared the numbers between minority and non-minority candidates for: (1) promotion rates; (2) passing rates, (3)

average exam scores; and (4) delays in promotion. 12/15/14 Tr. 81:23-82:11; Wiesen Report 3.

Dr. Wiesen, acknowledging that the first measurement, promotion rates, is the "most important[]," Wiesen Report 5, nevertheless argued that the other measurements were also important for various reasons. He argued that the passing rates and average scores were relevant because of the current system in which candidates are promoted in strict rank order. Id. He posited that passing rates and average scores were more "sensitive" or "statistical[ly] power[ful]" than promotion rates. Id. at 5-6. Dr. Wiesen opined that delays in promotion in the Department -- meaning the time between becoming eligible for a promotion and actually receiving that promotion -- were important because the timing of promotion from sergeant to lieutenant affects how job tasks are assigned. Id. at 6. Dr. Wiesen ultimately concluded that the 2005 and 2008 exams "had fairly severe adverse impact on minority candidates, black and Hispanic." 12/15/14 Tr. 48:21-24.

In contrast to the Plaintiffs' consideration of four sets of data, Dr. Silva and the City argued that promotion rates were the only appropriate measurement for determining adverse impact. Post-Trial City's Proposed Findings 44-45; see Silva Report 13. They argued essentially that the Court should ignore average exam scores and pass-fail rates because the only value they have

is in predicting promotion rates, which can measured directly. See Silva Report 13; Post-Trial City's Proposed Findings 45. Dr. Silva argued that a delay in promotion is an inappropriate measuring device because "the number of days between promotions is not a function of the test, it is a function of when the positions open up. The test is only responsible for the order in which the promotions are made." Silva Report 6. Analyzing only the promotion rates in 2005 and 2008 using a two-tailed test, the City argues, one cannot conclude that the 2008 exam resulted in a statistically significant adverse impact, and thus, judgment should enter in the City's favor. Silva Report 7-10; Post-Trial City's Proposed Findings 41.

The Court agrees with the Plaintiffs that statistics other than promotion rates are relevant in evaluating disparate impact. The City's argument that promotion rates are the only relevant factor is a "bottom line" defense which the Supreme Court has rejected. In the seminal case Connecticut v. Teal, the Supreme Court stated:

> In considering claims of disparate impact under [Title VII], this Court has consistently focused on employment and promotion requirements that create a discriminatory bar to opportunities. This Court has never read § 703(a)(2) as requiring the focus to be placed instead on the overall number of minority or female applicants actually hired or promoted.

Teal, 457 U.S. 440, 450 (1982). In Teal, a pass-fail test had an adverse impact on minorities but, due to an affirmative

action program, no adverse impact on promotion rates. Id. at 443-44. The Supreme Court rejected the agency's "bottom-line" defense, admonishing that "[t]he suggestion that disparate impact should be measured only at the bottom line ignores the fact that Title VII guarantees these individual respondents the opportunity to compete equally with white workers on the basis of job-related criteria." Id. at 451. In other words, "individual components of a hiring process may constitute separate and independent employment practices subject to Title VII even if the overall decision-making process does not disparately impact the ultimate employment decisions involving a protected group." Bradley, 443 F. Supp. 2d at 158-59. Under the progeny of Teal, even in the absence of adverse impact on promotion rates, an exam can lead to liability for an employer if it functions as "a gateway that has a disparate impact on minority hiring." Id. at 159. Promotion rates -- the "bottom line" -- are thus not the only relevant inquiry in this Court's disparate impact analysis.

The 2005 and 2008 exams served two functions: they were used as pass-fail hurdles and they accounted for 80% of the final score that determined candidates' rank on an eligibility list from which they were promoted in rank order. Under such a scheme, this Court cannot rule that passing rates and average scores are irrelevant. Indeed, the Second Circuit has ruled

that where an exam is used as both a pass-fail hurdle as well as a mechanism for ranking candidates for a promotion (as is the case here), courts should consider the disparate impact in the pass-fail rates, as well as the placement of ethnic groups on the ranking list. Waisome v. Port Auth. of New York & New Jersey, 948 F.2d 1370, 1377 (2d Cir. 1991). The average scores and passing rates are relevant to the Court's determination of whether the Plaintiffs have met their burden on prong one.[23]

The City argued that the timing of promotions is largely determined by the timing of vacancies, so delays are more relevant to the damages phase of the litigation than to the liability phase. See Post-Trial City's Proposed Findings 78. Neither argument persuades the Court to exclude delays in promotion from its analysis. The first argument falls flat because promotion rates themselves are determined, at least in part, by vacancies, and the City nowhere argues that promotion rates are irrelevant. Regarding the damages argument, the Court acknowledges that the delay in promotion will be relevant at the damages phase of the litigation, but it can also constitute disparate impact in the form of loss of pay, benefits, and seniority. See Bradley, 443 F. Supp. 2d at 168 (ranking by exam score disproportionately precluded minority candidates from

[23] In the City's own validation study of the 2002 exam, Morris & McDaniel based its statistical analysis on mean scores. Draft Validity Report 14; Assessment Center Report 18.

earning an earlier promotion, thus constituting an adverse impact); Guinyard v. City of New York, 800 F. Supp. 1083, 1089 (E.D.N.Y. 1992) (same). This comports with common sense, with case law, and with the spirit of Teal -- that employers may not circumvent Title VII "by merely showing that eventually they may hire some members of the disadvantaged minority group." Bridgeport Guardians, Inc. v. City of Bridgeport, 933 F.2d 1140, 1147-48 (2d Cir. 1991) (citing Teal, 457 U.S. at 455-56).

The Court will therefore consider all of the factors that Dr. Wiesen statistically analyzed: promotion rates, pass-fail rates, average scores, and delays in promotion.

2. To Aggregate or Not to Aggregate?

For two of the four aspects of the promotional procedure, promotional rates and passing rates, Dr. Wiesen aggregated the data for the 2005 and 2008 exams, Wiesen Report 10, 13, properly taking care to account for people who took both exams. Id. at 5. Dr. Wiesen stated that aggregation yields a "more powerful statistical test [because] you have a larger sample size and you see . . . the big picture." 12/15/14 Tr. 83:10-17. See also Wiesen Report 5.

Dr. Jacobs, a colleague of Dr. Silva's, argued in a pre-trial affidavit that aggregation was inappropriate. He posited that aggregation "only increases the sample size without a strong underlying logic as to the appropriateness of treating

candidates from 2005 and 2008 as competing for the same jobs." Aff. Rick R. Jacobs, PhD ("Jacobs Aff.") ¶ 15, ECF No. 93. Dr. Silva opined in his expert report that aggregating the data between 2005 and 2008 is inappropriate because of Simpson's Paradox, a phenomenon by which an "effect exists in two separate data sets but disappears when the data sets are combined or vice versa." Silva Report 10. During trial, Dr. Silva testified that aggregation creates a risk of distortion, even if Simpson's Paradox is not present. 12/18/14 Tr. 33:6-19, ECF No. 165.

Dr. Wiesen does not offer a sound basis for aggregating the data, other than its favoring the case of the party who hired him. He implies in his report that he only aggregated data if his original analysis did not produce statistically significant and practically important numbers for individual exam years. See Wiesen Report 16. In other words, he aggregated when the results using individual exams were not strong enough to help the Plaintiffs. The Court is not persuaded that this provides a sufficient basis to aggregate two data sets.[24]

This Court rules that aggregation is inappropriate in this case. The 2005 and 2008 exams presented different questions,

[24] In the Lopez case, Dr. Wiesen aggregated data across exam years and across jurisdictions within the Commonwealth. Lopez, at 20. Judge O'Toole found such aggregation inappropriate in part because of Simpson's Paradox, but also for other reasons. Id. at 18-20.

had different mean scores, and tested different candidates. 12/16/14 Tr. 129-131, ECF No. 162. Moreover, the Court has already ruled that the 2005 exam is not actionable; combining the 2005 scores with the 2008 scores would allow the Plaintiffs to circumvent this ruling. It would also raise so-called slippery slope concerns: why not aggregate with exams from the 1990s? Why not the 1980s? The Court sees no legitimate basis for aggregating the statistics on the facts of this case and therefore declines to do so.

3. One-Tailed vs. Two-Tailed

The next dispute the Court must resolve involves statistical methodology. It is a familiar one in disparate impact cases: whether to use a one- or two-tailed test when testing for statistical significance.

Drs. Wiesen and Silva both used "Fisher Exact Tests" to compare the exam results for candidates who are members of a minority group and white candidates. 12/17/14 Tr. 148:8-20. Fisher Exact Tests produce a bell curve with a tail on either end representing the lowest probability events. Id. at 149:24-150:9. When conducting a Fisher Exact Test, one can use a one-tailed or a two-tailed test. The terms "one-tailed" and "two-tailed" reflect whether statistical significance is determined from one or both the tails of the sampling distribution.

A two-tailed test assumes that any result could come from the test: in this case, in determining whether a given result was due to random chance, a two-tailed test would entertain three possibilities: that minorities would outperform non-minorities, non-minorities would outperform minorities, or that their performances would be equal. 12/17/14 Tr. 150:10-19. In contrast, a one-tailed test assumes only two possibilities: in this case, the performance between the groups was equal, or minorities performed worse on the test than non-minorities. Id. at 150:20-151:1; see Palmer v. Shultz, 815 F.2d 84, 94-95 (D.C. Cir. 1987).

In Dr. Wiesen's original report, he conducted all of his analyses using a two-tailed test, stating that although the one-tailed approach is more logically defensible, the two—tailed approach is more conservative. Wiesen Report 7 n.4; 12/16/14 Tr. 119:25-120:2. Subsequent to his initial report, and before the City offered its expert report, two additional sergeants were promoted to lieutenant, one white and one black. 12/15/14 Tr. 108:19-109:3.

When analyzing the data with the two new hires using a two-tailed test, Dr. Jacobs, the City's expert, concluded that the adverse impact for the promotion rates for the 2008 exam was not statistically significant. Jacobs Aff. ¶ 14. Dr. Jacobs found a p-value of .052, a hair above the .05 threshold. Id. ¶ 10.

Dr. Silva arrived at the same statistical conclusion. Silva Report 8.

Dr. Wiesen acknowledged this lack of statistical significance using a two-tailed test to examine the new data set. Second Aff. Joel P. Wiesen ("Wiesen Second Aff.") ¶ 6, ECF No. 103. He subsequently switched his analysis from a two-tailed test to a one-tailed test, defending this approach by explaining that the question asked in this litigation is whether there was adverse impact on minorities, and thus, the "one-tailed test is more appropriate." 12/16/14 Tr. 122; Wiesen Second Aff. ¶¶ 1, 7. Dr. Wiesen concluded that, even accounting for the two new hires, the "proper statistical conclusion is that there was adverse impact in promotions, both for the 2008 and the 2005 exams."[25] Id. ¶ 10.

The City bristled at Dr. Wiesen's flip-flop from the two-tailed test to the one-tailed test, arguing that the two-tailed test is the appropriate one because it "accepts it is possible on a promotional examination that minorities could outscore non-minorities or that non-minorities could outscore minorities[.]" Post-Trial City's Proposed Findings 41. Dr. Silva argued that a

[25] The p-value for the one-tailed test of the 2008 promotional rates was .027, which is less than .05 and is thus statistically significant. Wiesen Second Aff. ¶ 8. After running some additional analyses, Dr. Wiesen stated that the increase of the p-value above .05 using a two-tailed test was "a blip, not a pattern." Id. ¶ 6.

one-tailed approach was inappropriate because it "implies that there is no chance that the direction of a promotion rate difference will ever favor minorities," which contradicted Dr. Silva's professional experience. Silva Report 2-3.

Whether to use a one-tailed or a two-tailed test is a common point of contention in disparate impact cases. Title VII defendants often argue that both minority and majority groups are protected from discrimination, and "it is therefore inequitable to disregard the probability of outcomes that may favor either group." Palmer, 815 F.2d at 95. Defendants are also aware, of course, that it is easier for plaintiffs to prove significance and thus disparate impact with a one-tailed test. See, e.g., Brown v. Delta Airlines, Inc., 522 F. Supp. 1218, 1228 n.14 (S.D. Texas 1980).

The weight of the case law appears to favor two-tailed tests. In Palmer, the D.C. Circuit Court of Appeals favored the two-tailed test for Title VII cases, 815 F.2d at 95-96, and that court continues to do so today. Csicseri v. Bowsher, 862 F. Supp. 547, 564 (D.D.C. 1994) aff'd, 67 F.3d 972 (D.C. Cir. 1995). Other courts have agreed. See, e.g., Dicker v. Allstate Life Ins. Co., No. 89 C 4982, 1997 WL 182290, at *41 (N.D. Ill. Apr. 9, 1997) (two-tailed is especially appropriate in disparate impact claims involving facially neutral employment selection procedures).

Courts recognize, however, that a one-tailed test can be appropriate in some Title VII circumstances, such as when "one population is consistently over-selected over another." Stender v. Lucky Stores, Inc., 803 F. Supp. 259, 323 (N.D. Cal. 1992); see Brunet v. City of Columbus, 642 F. Supp. 1214, 1230 (S.D. Ohio 1986) rev'd on other grounds, 1 F.3d 390 (6th Cir. 1993) (stating that the one-tailed test is appropriate where "the raw numbers indicate that women are selected at a lesser rate than men. In these circumstances, the question being asked is whether this apparent difference is real or a statistical artifact."). A one-tailed test can also be appropriate if "[t]here is little chance, from a facial review of the evidence, that applicants" in the plaintiff class were "treated statistically better" than those in the other group. Csicseri, 862 F. Supp. at 564-65. The First Circuit has overtly avoided choosing between the two. Jones, 752 F.3d at 43 n.5.

There are two good arguments for using a one-tailed test in this case. The first is broader, relying on current inequalities in our society to put this case in context, and the second is narrower, implicating the history between this particular defendant, and this particular class of plaintiffs.

Experts for both sides testified to the phenomenon of written multiple-choice tests producing high levels of adverse impact on minority candidates. 12/15/14 Tr. 113-114; 12/18/14

Tr. 45-46, 70; Lopez, 07/13/10 Tr. 82-85; Lopez, 07/14/10 Tr. 43-48, 55, 59-60; Lopez, 07/26/10 Tr. 30; Lopez, 09/15/10 Tr. 58-59; Lopez, 09/16/10 Tr. 110. Judge O'Toole in Lopez also recognized this phenomenon. Lopez, slip op. at 14. Experts in the seminal case of Ricci v. DeStefano similarly testified. 557 U.S. at 570, 572 (2009). Why the difference in performance between members of minority groups and white applicants on these tests? Neither expert addressed this question, other than Dr. Wiesen suggesting that "there are . . . dozens of reasons, each of which account[] for just a small amount of that difference." 12/15/14 Tr. 114:10-13. Without wading into social-scientific debates, the Court agrees with Dr. Wiesen that there are likely several factors driving this disparity (e.g., legacies from historical discrimination, economic inequality, current explicit and implicit biases). Whatever the causes, the so-called "achievement gap" is real,[26] and might recommend adopting a one-tailed test as more rooted in reality.

[26] That it is an uncomfortable truth does not rob it of its current empirical foundation. There are many examples of the differences in academic achievement between white students and members of the minority groups at issue in this case (black and Hispanic).

This gap is present early. By fourth grade, white public school students outperform Hispanic students in both math and reading by the equivalent of roughly two grade levels. See U.S. Dep't of Educ., Nat'l Cent. for Educ. Statistics ("NCES"), How Hispanic and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress 10-11, 36-37 (June 2011) (describing point-

In addition, Boston has a history of discrimination against minority police officers. See Boston Police Superior Officers Fed'n v. City of Boston, 147 F.3d 13, 20 (1st Cir. 1998). This history might suggest that viewing white police-officer-applicants as equally likely to over- and under-perform minority applicants on an exam developed by Boston is overly idealistic.

The Court is hesitant, however, to analyze a disparate impact case, a case in which no one has accused the Department of any malfeasance or conscious desire to discriminate, under the assumption that minorities could only underperform and not also possibly outperform their white peers on a promotional exam. Cf. Parents Involved in Cmty. Sch. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 748 (2007) (Roberts, C.J.) (plurality opinion) ("The way to stop discrimination on the basis of race is to stop discriminating on the basis of race."); id. at 789 (Kennedy, J., concurring) (suggesting that "facially race-neutral means" or considering race as part of "a more nuanced,

---

differential on tests; providing scores for grades four and eight for estimating grade-level gains). The same goes for white public school forth-graders as compared with black fourth-graders. See NCES, How Black and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress iii (July 2009).

The gap also manifests itself in later measures of academic success. For one, the four-year high school graduation rate for white students is currently 84%, for black students 69%, and for Hispanic students, 71%. NCES, Public High School Four-Year On-Time Graduation Rates and Event Dropout Rates: School Years 2010-2011 and 2011-12 4 (April 2014).

individual" evaluation of applicants is permissible, whereas broad-based classifications based on it are subject to strict scrutiny).[27] Moreover, although the First Circuit and the Supreme Court have remained relatively quiet on the issue of one versus two tails, the Supreme Court has previously suggested that a protected class's treatment that falls two or three standard deviations beyond the mean is evidence of disparate impact. See Castaneda v. Partida, 430 U.S. 482, 496 n.17 (1977). This requirement is closer to that of a two-tailed test (which requires a result be more than 1.96 standard deviations from the mean to reach statistical significance) than of a one-tailed test (1.65). The Court is inclined to agree with the City that a two-tailed test is the more appropriate methodology for evaluating statistical significance in this case.

The debate over the one versus two-tailed test, while important and fascinating, is not dispositive in this case. As the Court explains in the next section, the Plaintiffs have met

[27] Parents Involved, of course, has been the subject of considerable scholarship, with the Chief Justice's quip, quoted above, raising particularly strong objections, see, e.g., James E. Ryan, The Supreme Court and Voluntary Integration, 121 Harv. L. Rev. 131, 156-57 (2007) (characterizing the pronouncement as saying "we know how to end race discrimination in this country and you [a Seattle school district, in that case,] who are closer to the issue than we will ever be and have been working in good faith toward the same end, do not.") (internal footnote omitted). The Court does not comment on the merits of the Supreme Court's analysis in Parents Involved but simply notes that applying a one-tailed test would appear to be in tension with its reasoning.

their burden of demonstrating disparate impact, regardless of whether the Court accepts the one or two-tailed approach for the 2008 promotion rates.

4. Conclusions Regarding Disparate Impact (Prong 1)

In his expert report, Dr. Wiesen concluded that:

- The adverse impact ratio for promotions from the 2008 exam was .45 (meaning that minorities were promoted at a rate of .45 as compared to the rate at which whites were promoted, well below the 4/5 rule, which would be satisfied by a rate of up to .80), Wiesen Second Aff. 7, and .38 for the 2005 exam. Wiesen Report 8. He concluded that the p-value for the 2008 promotion rates was .052 for a two-tailed test, and .027 for a one-tailed test. Wiesen Second Aff. 7. The ratio for the 2005 exam was not statistically significant. Wiesen Report 9.

- The adverse impact ratio for passing scores was .74 for the 2008 exam and .57 for the 2005 exam, both of again satisfy the 4/5 rule, and both of which were statistically significant at a p-value of .004 for the 2008 exam and .00005 for the 2005 exam, using a two-tailed test. Wiesen Report 11-12; 12/15/14 Tr. 88:1-4.

- The adverse impact for average scores was 6.6 points on the 2008 exam (meaning that minority applicants scored, on average, 6.6 points lower than white applicants) and 8.8 points on the 2005 exam, both of which were "highly statistically significant" at a p-value for the 2008 exam of .0015 and for the 2005 exam of .00003 (both using a two-tailed test). Wiesen Report 3, 15-16.[28]

- The adverse impact for delay in promotions for the 2008 exam was an average of 750 additional days, which was statistically significant at a p-value of .001 (using a two-tailed test). Wiesen Second Aff. 4-5.

[28] This analysis was based on the scores from the multiple choice exam alone. Id. at 13.

> There was no adverse impact for promotion dates from the 2005 exam. Wiesen Report 16-17.

The Court concludes that the Plaintiffs have met their burden of establishing disparate impact stemming from the 2008 exam. The fact that the p-value for the 2008 promotion rates was .052 using a two-tailed test, a breath above the .05 threshold, is not enough to persuade the Court that the Plaintiffs failed to meet their burden of establishing a prima facie case of disparate impact. This is so for various reasons:

First, while the .05 cut-off for a finding of statistical significance is generally accepted, such thresholds are not hard-line rules in disparate impact cases. E.g., Jones, 752, F.3d at 46-47 (explicitly declining to rule on whether establishing a p-value of .05 was a requirement to show disparate impact); Little v. Master-Bilt Products, Inc., 506 F. Supp. 319, 333 (N.D. Miss. 1980) (rejecting a hard-line cut-off of a certain p-value to make out a prima facie Title VII case based on statistical evidence alone).

Second, the City's expert on test validity acknowledged that when significance is between .05 and .10, the results are "marginally significant." 01/05/2015 Tr. 53:2-15. He further acknowledged that a marginally significant analysis, combined with other analyses that are statistically significant, can lead to a finding that an adverse impact has been demonstrated. Id.

As explained above, the Plaintiffs have presented evidence of statistically significant disparate impact on pass-fail rates, averages scores, and delay in promotions.

Third, the First Circuit has acknowledged that when evaluating statistical significance, "no single test controls in measuring disparate impact." Langlois, 207 F.3d at 50 (citing Watson, 487 U.S. at 995-96 n.3) (noting that a case-by-case approach is appropriate); see also Waisome, 948 F.2d at 1376 (noting that courts should "consider[] not only statistics but also all the surrounding facts and circumstances"); Police Officers for Equal Rights v. City of Columbus, 644 F. Supp. 393, 432 n.13 (S.D. Ohio 1985) (stating that even if plaintiff fell short of cut-offs for statistical significance, "other evidence submitted in this case . . . tends to support the inference of disparate impact suggested by plaintiff's statistical data"). The Plaintiffs bolster their evidence on statistical significance by presenting evidence that the 2005 and 2008 exams violated the four-fifths rule.[29] As this Court has already held,

[29] Using Monte Carlo simulations, Dr. Silva concluded that Dr. Wiesen's adverse impact ratio calculations (which form the basis of the four-fifths rule) had a high error rate, and the Court should therefore disregard them. Silva Report 5, 10. The Court might find this argument more persuasive if the only evidence Plaintiffs had put forth was a violation of the four-fifths rule. Such is not the case here. Moreover, the City points to no case law addressing a statistical analysis of error rates for the four-fifths rule, and the Court sees no obvious reason for such an analysis. The four-fifths rule, unlike a

"a violation of the four-fifths rule . . . may demonstrate adverse impact, particularly when coupled with other statistical evidence of adverse impact." Cotter, 193 F. Supp. 2d at 348 n.12; see also Bradley, 443 F. Supp. 2d at 163 (holding that the plaintiffs established a prima facie case of statistical significance by using the four-fifths rule combined with a chi-square analysis).[30]

The Court also considers the fact that if the eligibility list from the 2008 exam had not been extended due to the Lopez litigation, but instead had expired three years after its

---

measurement of statistical significance, is a rough and ready rule of guidance which is not given decisive weight. It is a "rule of thumb" only, and the Court regards it as such.

[30] Both parties, citing to Jones, agree that the four-fifths rule cannot, by itself, be given decisive weight. 12/15/14 Tr. 47:2-14. In Jones, the First Circuit held that defendants in disparate impact cases could not use the four-fifths rule to trump a statistically significant showing of disparate impact. Jones, 752 F.3d at 52. The City cites to this case law in an effort to convince this Court to disregard the four-fifths rule altogether. 12/15/14 Tr. at 46-47; Def.'s Mot. In Limine Exclude Use "Four-Fifths Rule" Prove Disprove Adverse Impact Based Race, ECF No. 131. The First Circuit did not hold that the four-fifths rule has no place in disparate impact jurisprudence; the First Circuit specifically noted that the regulation establishing the four-fifths rule shows "that the commission views practical significance, along with statistical significance, as relevant in identifying a disparate impact." Jones, 752 F.3d at 50 (citing 29 C.F.R. § 1607.4(D)). The Court went on to say that the "four-fifths rule may serve important needs," is a helpful "rule of thumb," and "has some practical utility." Jones, 752 F.3d at 52. The Supreme Court cited to the four-fifths rule in the very case to which the City so extensively cites in its post-trial briefing. Ricci, 557 U.S. at 586-87; Post-Trial City's Proposed Findings 4.

creation, as is typical, not a single black sergeant would have been promoted to lieutenant. 01/06/15 Tr. 144-45. Lastly, the Court views as likely the probability that the 2008 figures underestimate the disparate impact considering that at least some of these test-takers had presumably passed the 2005 sergeants' exam, which Boston conceded as having a disparate impact on minority candidates. Lopez, at 20; see also Nash v. Consol. City of Jacksonville, Duval Cty., Fla., 895 F. Supp. 1536, 1544 (M.D. Fla. 1995) aff'd sub nom., 85 F.3d 643 (11th Cir. 1996) (holding that the plaintiff made out a prima facie case of disparate impact despite a lack of statistical significance based on past discrimination which reduced the number of African-Americans eligible to sit for the promotional exam).

All of this evidence combined is enough for this Court to rule that the Plaintiffs have met their burden of raising an inference of causation and demonstrating a prima facie case of disparate impact. The burden thus shifts to the City to defend its promotional process as a valid selection tool.

B. Job-Related and Consistent with Business Necessity (Prong 2)

1. Introduction

To pass muster under Title VII once disparate impact has been shown, the City must convince the Court that the 2008 exam

was both "job related" for the position of Boston Police Department lieutenant and consistent with "business necessity." Jones, 752 F.3d at 53(quoting 42 U.S.C. § 2000e–2(k)(1)(A)(i)). The purpose of this second prong is not for the Court to substitute its judgment for the City's or for that of experienced police officers, but to ensure that the City took sufficient care to ensure that its employment practice was consistent with business necessity. Cf. Texas Dep't of Hous., 135 S. Ct. at 2512 (holding that "[p]olicies, whether governmental or private, are not contrary to the disparate-impact requirement unless they are artificial, arbitrary, and unnecessary barriers") (internal quotation marks omitted); Langlois, 207 F.3d at 54 (Stahl, J., dissenting) (stating, in the discriminatory housing context, that a practice fails a similar standard if, "without demonstrably advancing the interest asserted in justification, [it] somehow impedes persons of color from competing on an equal footing with others").

The First Circuit has instructed that defendants whose employment practices produce a disparate impact must establish two elements to satisfy this second prong: "First, the [defendant] must show that its program aims to measure a characteristic that constitutes an 'important element of work behavior.'" Jones, 752 F.3d at 54 (quoting Albemarle Paper Co. v. Moody, 422 U.S. 405, 431 (1975)). Second, the defendant

"must show that the outcomes of [its challenged practice] are 'predictive of or significantly correlated with' the characteristic described above." Id. (quoting Albermarle Paper Co., 422 U.S. at 431). This framework directly mirrors that in the Uniform Guidelines, compare 29 C.F.R. § 1607.5(B) ("[T]he selection procedure [must be] predictive of or significantly correlated with important elements of job performance.").

The Guidelines provide a sensible way of evaluating whether a given test, like the one in this case, measures an important work characteristic, and whether the outcomes of that test are actually correlated with the characteristic measured. Many courts have utilized the Guidelines in exactly this way, and have re-phrased the prong two inquiry (following the Guidelines' language) as a determination of whether a given test is "valid." See, e.g., M.O.C.H.A. Soc'y, Inc. v. City of Buffalo, 689 F.3d 263, 274 (2d Cir. 2012) (discussing validity as goal of prong 2 inquiry; utilizing the Uniform Guidelines in its analysis); Bryant v. City of Chicago, 200 F.3d 1092, 1094 (7th Cir. 2000) (same); Williams v. Ford Motor Co., 187 F.3d 533, 539 (6th Cir. 1999) (same); Lopez, at 28 (same); Bradley v. City of Lynn, 443 F. Supp. 2d 145, 161 (D. Mass. 2006) (Saris, J.) (same). This Court follows their lead, and will look to the Uniform Guidelines throughout its prong 2 analysis.

2. Did the Department Aim to Measure a Characteristic that Constitutes an Important Element of Work Behavior?

Beginning the First Circuit's two-part inquiry for prong two, the Court must determine whether the 2008 exam measures "characteristic[s] that constitute[] . . . important element[s] of work behavior." Jones, 752 F.3d at 54 (internal citation omitted). In short, the Court holds that it does.

The development of most promotional examinations begins with a job analysis. A job analysis "includes an analysis of the important work behavior(s) required for successful performance and their relative importance." 29 C.F.R. § 1607.14(C)(2). It need not be developed concurrently with an exam to form the basis of the exam's validation. See Rudder v. District of Columbia, 890 F. Supp. 23, 42 (D.D.C. 1995) (a job analysis created earlier which is itself based on an even older job analysis retains relevance when officials were interviewed to ensure the job analysis was still relevant and the job had not changed significantly). A job analysis typically involves Subject Matter Experts (again, SMEs) identifying important tasks for the job, followed by identifying Knowledge, Skills, and Abilities (again, KSAs) necessary to perform those tasks, followed by linking the KSAs back to the tasks. E.g., United States v. City of New York, 731 F. Supp. 2d 291, 302 (E.D.N.Y. 2010).

Dr. Campion testified that the 1991 and 2000 job analyses, on which the 2008 exam was based, were conducted in great detail and reflected good linkage between KSAs and tasks. 12/19/14 Tr. 32:13-16, 36-37. He testified that the age of the 1991 job analysis did not undercut its usefulness, id. at 15:3-23, and the 2005 and 2008 mini-job analyses, while not as thorough as the 1991 and 2000 job analyses, were nonetheless sufficient because the job of lieutenant had not changed much over time. Id. at 67:14-68:1; Campion Report 21-22.

Dr. Wiesen testified that the 2000 job analysis –– which was the focus of his report -- was insufficient. He opined that the SMEs' rankings were not done with sufficient care or attention, reflected by the fact that the SMEs unanimously agreed that they performed the tasks of reporting homicides and disciplining subordinates on a daily basis, which, based on the statistics, is impossible. See Wiesen Rebuttal 15-16. He noted that SME ratings for the second half of the list of 302 tasks became much more uniform and cursory, likely reflecting rater fatigue. Id. at 87. He complained that mis-numbering of tasks and KSAs likely led to inaccurate ratings. Id. at 88. He also complained about the wording of the tasks in the final task list. Id. at 89.

While acknowledging that the 2000 job analysis may have some errors and is not a model of perfection, the Court

concludes that the robust job analyses performed in 1991 and 2000, and the mini-job analyses performed in 2005 and 2008, were adequate. As detailed above, for the 1991 job analysis, DPA identified important work behaviors by gathering information from various sources, created a list of tasks, asking police officers to rate the tasks, creating a list of potentially important KSAs, and asking SMEs to link KSAs to tasks. The same was true for the 2000 job analysis for which Morris & McDaniel, in part using the 1991 report, created a list of 302 possibly relevant tasks and KSAs, which SMEs rated. In 2005 and 2008, HRD asked SMEs to again rank tasks and KSAs that had been identified in the older reports.

Although there may have been some errors or snags in the review system –– for instance, there seems to be no evidence that the SMEs in 2008 linked the important KSAs back to tasks –– the lists of important tasks and KSAs emerging from the various job analyses strike the Court as adequately capturing the role of a Boston Police lieutenant. Dr. Wiesen nowhere claims that the lists of important tasks and KSAs have glaring gaps, and Plaintiffs' expert Dr. Hough opined that the 145 KSAs identified in the 2000 job analysis were consistent with the studies she had done in the field. 01/06/15 Tr. 19:5-10. The Court finds that the job analyses were sufficiently thorough and current so as to form solid ground on which to build a valid test; in other

words, they measure "an important element of work behavior." Jones, 752 F.3d at 54 (internal citation omitted). The City therefore has satisfied its burden for the first element of prong two.

3. Were the Exam Results Predictive of or Correlated with the Important Work Behaviors?

Dr. Michael A. Campion, who holds a PhD in industrial psychology, served as the City's expert on the issue of validity. 12/19/14 Tr. 5:21-6:4. In his expert opinion, the 2008 exam was content valid (more on that below) because it complied with the Uniform Guidelines "well enough," 12/19/14 Tr. 54:22-55:2, complied with the Society for Industrial Organizational Principles, Ex. 73, Report Michael A. Campion, PhD ("Campion Report") 21, and conformed with best practices in test development, id. at 25.

Dr. Wiesen again served as the Plaintiffs' expert to discuss validation. He concluded that the 2008 exam was not content valid. In doing so, he criticized nearly every step of the City's validation process, including its job analyses, Wiesen Rebuttal 15, the construction of the exam, id. at 18, the content of the exam, id. at 26, and how the exam was used (here, to rank candidates), id. at 28-32. Although the Court is not persuaded by all of Dr. Wiesen's criticisms, the Court ultimately agrees with Dr. Wiesen that the evidence does not

support the necessary inference that those who perform better on the exam will be better performers on the job, primarily because the exam did not test a sufficient range of KSAs, and there was no evidence that the exam was reliable enough to justify its use for rank ordering.

a. Methodology

Employers are not required to submit formal validation studies to meet this standard, but neither are courts required to take a leap of faith and simply accept an employer's claim to validity. Bradley, 443 F. Supp. 2d at 171 (citing Watson, 487 U.S. at 998; Steamship Clerks, 48 F.3d at 607). Validation is not primarily a legal subject, and requires expertise that courts lack. Guardians Ass'n of New York City Police Dep't, Inc. v. Civil Serv. Comm'n of City of New York, 630 F.2d 79, 89 (2d Cir. 1980).

There are two methods commonly used for determining validity: content validity and criterion validity. 29 C.F.R. § 1607.5(B). Criterion validity is a statistical analysis in which an analyst correlates the selection procedure with job performance. 12/19/14 Tr. 19:25-20:21. Content validity, by contrast, is not statistical; it is an attempt to link the important KSAs of the job with the selection procedure. Id. at 21:3-10. Under either approach, the goal is to make inferences from the test scores about future job performance. Id. at

22:10-16. Here, the City attempted to shoulder its burden on job relatedness using content validity. Id. at 25:14-19; Campion Report 5-6; Wiesen Rebuttal 8.

In evaluating content validity, experts in the field[31] rely on three authorities: the Uniform Guidelines, mentioned above, the "Standards for Educational Psychological Tests," referred to as "the Standards," and the "Society for Industrial

[31] Fact witnesses for both sides offered their opinions regarding the ability of the Department's promotional procedures to predict job performance. Commissioner William Evans testified that he had observed a direct relationship between how well a person did on a Department knowledge exam and how well they performed on the job. 01/07/15 Tr. 41-42, 46. Commissioner Evans' predecessor, Commissioner Davis, disagreed. He thought that communication and interpersonal skills are very important for a superior officer, 01/05/15 Tr. 91:3-11, and the inclusion of assessment centers (in addition to knowledge tests) improved the Department's promotional procedures. Id. at 91:17-92:1. In his opinion, past job performance is likely one of the best indicators for core lieutenant skills. Id. at 113:5-15. Commissioner Davis's predecessor, Commissioner Paul Evans (coincidentally, the brother of the current Commissioner) expressed his opinion in a memo that "the best indicator of future performance is past performance . . . . The best supervisors cannot always be identified solely by their performance on a written test and an hour in an assessment center. . . . We must become willing to reward police work, not memorization skills." Lopez, Ex. 194.

The named plaintiff in this case, Bruce Smith, testified that the 2008 exam did not ask questions that would differentiate his ability to serve as a lieutenant as opposed to a sergeant. 12/17/14 Tr. 109:7-15. Although the Court regards the fact witness' testimony as important for certain aspects of this case, and cites to it accordingly, the fact witnesses' personal opinions on the effectiveness of the testing procedures do not aid the Court on this prong of the framework, which requires the Court to rely on testing expertise. Guardians, 630 F.2d at 89.

Organizational Psychology" principles, referred to as "the Principles." 12/15/14 Tr. 118:4-19; Ex. 49, Rebuttal Dr. Campion's Report Validity Alternatives 2008 and 2005 Exams ("Wiesen Rebuttal") 10.

b. The Scope of the Analysis

The 2008 exam consisted of two components: the written, multiple-choice component (weighted at 80%), and the E&E component (weighted at 20%). 12/15/14 Tr. 49:25-50:17. In analyzing the validity of the 2008 exam, the Court will exclude the E&E portion for the reasons detailed below.

Every candidate that sits for the exam is automatically awarded fourteen of the twenty possible E&E points. Id. at 50:24-51:5. Dr. Wiesen posits that, because of this practice, the E&E portion in actuality has very little bearing on where the candidate ranks on the eligibility list. Id. In his report, Dr. Wiesen correlated candidates' scores on the written exam with their final exam score, calculating a correlation coefficient of .95, an almost perfect positive correlation. Wiesen Report 20.[32] In other words, the E&E component and the bonus points awarded for things such as veteran's status had virtually "no impact on the final exam scores." 12/15/14 Tr. 58:19-21. It is the score on the written exam, Dr. Wiesen

[32] For the 2005 exam, the coefficient was .96. Wiesen Report 21.

argues, that drives a candidate's placement on the eligibility list.

The City counters this point by arguing that the fourteen points automatically awarded to every candidate should not be discounted because they measure characteristics important to serving as a lieutenant. By way of analogy, the City offers that medical degrees, while held by every doctor, are made no less important in a promotional context simply because every doctor holds one. Post-Trial City's Proposed Findings 58-59; 01/05/15 Tr. at 56:11-20.

The City's argument is not persuasive. The Court does not deny that the length of employment and nature of experience of a sergeant in the Department, much like a doctor's medical degree, is an important input to a candidate's performance on the job. But the medical degree, while an important requirement for a doctor, would not aid the hospital in selecting which doctors to promote. Dr. Campion acknowledged as much when he testified in the context of this doctor analogy that a medical degree "may not differentiate [the doctor candidates] very much." 01/05/15 Tr. at 56:12-20. Here, the Court is evaluating the Department's method for selecting whom to promote within a pool of candidates who all have earned fourteen points. In short, differentiation is the crux of this case. Dr. Wiesen's analysis persuades this Court to find that the E&E portion of the exam had a de minimis

impact on the candidates' final scores. The Court will therefore focus on whether the 2008 written exam was content valid.

In doing so, the Court will evaluate the adequacy of the test construction process used by the City, the exam's content (the extent to which it matches up with the job), and how the Department used the exam (meaning its system of scoring the exam and then using those scores to rank candidates). Cf. Guardians, 630 F.2d at 95 ("distill[ing] five attributes" from the Guidelines relevant to assessing content validity covering "the quality of the test's development[,]" the test's content, and the test's grading).

c. Adequacy of Test Construction

Having explained the overall framework and appropriately narrowed the scope of its inquiry, the Court finally begins its analysis of the second element of prong two, looking first to test construction. The Court finds that the City fell short at many stages of the test construction process. In his expert report, Dr. Campion addresses many factors that go into test development. The Court will address those that strike it as most relevant to deciding the instant case.

When using a multiple choice exam, the developer must convert the job analysis result into a test plan to "ensure a direct and strong relationship between the job analysis and the

exam." Campion Report 24. As Dr. Campion points out, the City created test plans for the 2008 exam. Ex. 54; Ex. 60. The Court cannot find, however, that the test plan ensured a strong relationship between the job analysis and the exam. As discussed above, the test outlines reflect that the 2008 exam was written to test knowledge; the outlines reveal that only two abilities appeared on the 2008 exam. Because the Court has found that the E&E component had very little bearing on the final score, the Court cannot find that the City met its burden on this step: too many skills and abilities were missing from the 2008 test outline.

Next, an employer should use validation analyses through a variety of measures, including "ratings of test items, [and] linkages between test content and job tasks or KSAs." Campion Report 24. HRD conducted a robust validation analysis in 1991 after administering the 1991 exam. As part of this process, HRD asked SMEs to again link tasks with KSAs, link tasks with examination subjects, and link KSAs with examination subjects, and evaluate the manner in which each was covered on the 1991 written test. 1991 Validation Report at 00350. While the process was not as robust in 2008, it does appear as though the SMEs reviewed the test questions, identified which KSAs matched

the questions,[33] and evaluated the questions for difficulty, readability, and recommendation for use. Ex. 60. The City has met its burden on this requirement.

The next step is for the test developer to "conduct[] various statistical analyses to ensure the quality of the test scores," such as item analyses, reliability analyses, and adverse impact analyses. Campion Report 24-25. The record does not present evidence that HRD conducted such analyses following the 2008 exam. Dr. Campion cites exclusively to the 1991 report to demonstrate that the City conducted the necessary analyses to ensure the quality of the test scores. Campion Report 25. Considering the updates to the tasks and KSAs in 2000, 2005 and 2008, and the lack of evidence in the record that the 2008 exam was highly similar to the 1991 exam, the Court finds this citation insufficient. The Court thus finds that the City failed to conduct statistical analyses to ensure the quality of the test scores for the 2008 exam.

Lastly, the test developer must recommend proper administration of the test, including cut-off scores, rankings,

[33] The Court notes that it is very difficult to tell from the exhibits exactly what the SMEs did in 2008 to support validity. Dr. Wiesen complained that HRD insufficiently explained how it prepared the 2008 exam. Wiesen Rebuttal 18. This lack of explanation, he asserted, flew in the face of published professional standards for testing. Id. The City would do well to improve its procedures for documenting its validation processes.

bands, and weighting. Campion Report 25. There was no indication in the record that HRD analyzed any of these things in 2008. Dr. Campion, again, points to the 1991 report to support his conclusion that HRD fulfilled its obligation to properly construct the test. Id. Again, the Court is troubled by this because the 2008 and 1991 exams were different.

Further, the Court is troubled by the substance of the 1991 report on these points. For instance, to explain its decision in 1991 to weight the written portion of the exam at 80% and the E&E at 20%, HRD looked to the 1985 and 1987 sergeant promotional exams, for which SMEs recommended weights for the various components. 1991 Validation Report at 00349. Those exams included an interactive component in addition to the written component. For the 1991 exam, which did not include the interactive component, HRD chose to weight the written portion at 80%, stating that "if the same SMEs were asked to split the full examination between only two components [as opposed to three components] . . . the education and experience component would receive more weight than previously." See 1991 Validation Report at 00349. This explanation cannot alone support the City's decision, seventeen years later, to weight the written portion of the exam at 80%. The Court also fails to see how the 1991 report supports the 2008 cut-off score of 70. The Uniform Guidelines state that "[w]here cutoff scores are used, they

should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force." 29 C.F.R. § 1607.5(H). The only criteria HRD seemed to use in 1991 to set a cut-off score was adverse impact; there is no mention of proficiency. 1991 Validation Report at 00381. The City has not explained why it chose 70 as a cut-off score for the 2008 exam.

Considering all of these factors,[34] the Court finds that the test construction process was inadequate to support the heightened validity requirement necessary to rank candidates.[35]

[34] Dr. Campion also argued that, when developing a test, "[t]he test developer should evaluate alternative test methods that would most validly measure the KSAs needed at time of hire; if this is in the employment context, this will also usually consider the potential adverse impact of the test method." Campion Report 23. There is indeed no evidence in the record that the City in 2008 evaluated alternative test methods. In defense of the City, Dr. Campion cites to HRD's conclusion in the 1991 report that "(1) job knowledge tests have been shown by the research to be generally valid for a very wide variety of jobs, (2) reading and writing are required on the job so a written test is a reasonable format, and (3) a multiple-choice format is feasible to administer to a large applicant population." Campion Report 23. If the City did in fact have such an obligation to consider alternatives in 2008 (an issue the Court need not decide, as its prior analysis of factors is sufficient to decide the issue), the City's use of this prior decision, which did not explicitly consider relative validity, would not excuse that obligation.

[35] For a discussion of the heightened standards for scores that rank candidates, see infra Part V-B-3-e.

d. Content of the Exam

The Court, proceeding with its own validity analysis,[36] continues to rely on expert testimony and the Uniform Guidelines to structure its inquiry, which now shifts to the content of the 2008 exam.

The Uniform Guidelines state that "[a] selection procedure can be supported by a content validity strategy to the extent that it is a representative sample of the content of the job." 29 C.F.R. § 1607.14(C)(1). If the selection procedure purports to measure knowledge, skills, and abilities -– as is the case here –- the employer "should show that (a) the selection procedure measures and is a representative sample of that knowledge, skill, or ability." Id. § 1607.14(C)(4).

The Second Circuit has stated that the purpose of this requirement is

> to prevent either the use of some minor aspect of the job as the basis for the selection procedure or the needless elimination of some significant part of the job's requirements from the selection process entirely; . . . . Thus, it is reasonable to insist that the test measure important aspects of the job, at least those for which appropriate measurement is

[36] The Court does so because the City did not conduct a formal validity study of the 2008 exam. The City cannot establish validity by relying on the 1991 validation study; the study was conducted seventeen years before the 2008 exam and experts generally agree that validity studies should be conducted every five to eight years. Bradley, 443 F. Supp. 2d at 172. Seventeen years "is too long a hiatus under the standards in the industry." Id. Further, the 1991 validity study did not validate the 1991 exam for strict rank order.

> feasible, but not that it measure all aspects, regardless of significance, in their exact proportions.

Guardians, 630 F.2d at 99.

From the 1991 job analysis, DPA determined that 187 KSAs were necessary to carry out the tasks critical to performing the role of a Boston police lieutenant. 1991 Validation Report at 00257-58. Of these 187 KSAs, 108 were skills and abilities. 1991 Validation Report, App. EE. The 2000 report concluded that 145 KSAs were critical to the role. 2000 Job Analysis 49. Of these 145 critical KSAs, 91 were skills and abilities. Id. at 59-64.

The Court finds that the thirteen knowledge categories were very broadly worded (much more so than the knowledge categories emerging from the 1990 and 2000 reports), and is therefore satisfied that the 2008 exam tested a sufficient range of the critical knowledge areas.[37] Yet that is not enough to demonstrate content validity: the 2008 test outline indicated that HRD decided to test none of the critical skills, and to test only two abilities categories: the ability to read, understand, interpret, and explain material in written form, and

[37] The test outline for the 2008 exam reflects that HRD decided to test only thirteen knowledge categories. Ex. 54; Ex. 60. The categories were, however, worded more broadly than those appearing on the 1991 and 2000 job analyses. In fact, Dr. Campion opined that the knowledge categories were so broad that roughly 80% of the knowledge areas from the 2000 report could fall under these thirteen categories. 01/05/15 Tr. 60:5-16.

the ability to read and interpret documents such as maps and charts, and make basic arithmetical calculations. Ex. 54; Ex. 60.

The Court concludes that the 2008 exam did not sufficiently test for a representative sample of the critical KSAs. The job analyses reflect that many skills and abilities are necessary to perform the job of lieutenant, yet the 2008 written exam tested knowledge almost exclusively. Dr. Campion's report does little to convince the Court otherwise. He stated that the examinations "are representative measures of the knowledge areas" (but no mention of the skills or abilities), Campion Report 10, and that the exam "measured a large number of knowledge areas," (again, without mention to the skills and abilities), id. at 14. His report discussed at some length how well knowledge was tested, again, with no mention of the skills or abilities. Id. at 14-15. Dr. Campion in fact acknowledged during trial that there was no attempt to test for skills and abilities that may be important to a lieutenant, 12/22/14 Tr. 32:2-9. Despite these deficiencies, he opined that the work behaviors selected for measurement were important and constituted most of the job. Campion Report 13.

Because job knowledge is only a limited part of the job analyses for the role of lieutenant, the Court agrees with Dr. Wiesen that the 2008 exam skipped over critical skills and

abilities, including interpersonal skills, presentation skills, reasoning and judgment skills, oral communication skills, analytical skills, ability to give constructive criticism, ability to speak in front of groups, ability to counsel subordinates, ability to counsel and comfort families of victims, and ability to make sound decisions quickly. Wiesen Rebuttal 28. The Court therefore also agrees with Dr. Wiesen's opinion that, as a result of HRD's decision not to test many critical skills, a high score on the 2008 exam simply was not a good indicator that a candidate would be a good lieutenant. Id. While the Court acknowledges that an exam need not test for every relevant KSA to be content valid, and gives credence to Commissioner Evans' statement that "knowledge is power," 01/07/15 Tr. at 20,[38] the near total absence of the 2008 exam's test of critical skills and abilities leads the Court to conclude that the City has not demonstrated validity.[39] This is

[38] Commissioner Evans also testified that in his experience, police officers who scored higher on the written exam were better performers. 01/07/15 Tr. 41:18-20. Such anecdotal accounts, however, "cannot substitute for actual evidence of validity." United States v. City of New York, 637 F. Supp. 2d 77, 131 (E.D.N.Y. 2009). In any event, former Department Commissioners evidently placed less weight on the written exam. See supra notes 16, 31 (advocating for an assessment center).

[39] The Court also notes that the City nowhere substantiated its use of 70 as a cut-off score. There should be an independent basis for choosing a cutoff score, such as a determination that the cut-off point separates those who can do the job from those who cannot, or a way of locating a logical

consistent with Judge O'Toole's opinion in Lopez in which he held that the written portion of the 2008 sergeants' exam could not alone support 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." Lopez, at 35-36;[40] cf. Boston Police Superior Officers Fed'n v. Civil Serv. Comm'n, 35 Mass. App. Ct. 688, 695 (1993) (ruling that the Civil Service Commission properly concluded that the 1987 Boston lieutenants' exam, which consisted of a multiple-choice written exam and an E&E component after other components were thrown out, "failed to constitute a fair test of supervisory skills and ability").

e. Use of Exam Results to Rank Candidates

The Court analyzes how the exam results were used because "evidence of both the validity and utility of a selection procedure should support the method the user chooses for operational use of the procedure, if that method of use has a

---

break-point in the pool of candidates. United States v. City of New York, 637 F. Supp. 2d 77, 124 (E.D.N.Y. 2009).

[40] Dr. Campion points out that in the 2000 report, SMEs rated the extent to which all of the KSAs differentiated performers, and only those KSAs that differentiated between good and bad performers were used to develop the exam. Campion Report 15-16. The Court agrees with Dr. Campion that this would support the City's ability to distinguish among candidates' command of the requisite knowledge areas, but it does not address the utter lack of skills and abilities tested on the 2008 exam.

greater adverse impact than another method of use." 29 C.F.R. § 1607.5(G). Even were the Court able to find that the 2008 exam was sufficiently valid as a screening tool, the Court would still hold the 2008 exam invalid as the sole mechanism for ranking candidates.

The Uniform Guidelines impose a higher standard for finding validity when a test is used to rank candidates as compared to when a test is used simply as a screening tool, stating that "[e]vidence which may be sufficient to support the use of a selection procedure on a pass/fail (screening) basis may be insufficient to support the use of the same procedure on a ranking basis." Id.; cf. Boston Chapter, N.A.A.C.P., Inc. v. Beecher, 504 F.2d 1017, 1026 (1st Cir. 1974) ("Even if the test is minimally valid, we might also doubt its use as an absolute cutoff. A test seemingly should receive no more weight in the selection process than its validity warrants. Use of a minimally valid test as an absolute cutoff is questionable even if more limited uses of the test are acceptable.").

The Uniform Guidelines specifically allow employers to demonstrate validity for rank-order exams using content validity. To do so, the employer must show "by a job analysis or otherwise, that a higher score on a content valid selection procedure is likely to result in better job performance." Id. § 1607.14(C)(9). Although the Court may infer a relationship

between higher test performance and higher job performance, when the test scores reveal adverse impact that is greater at the higher end of the ranked list, "the appropriateness of inferring that higher scores closely correlated with better job performance must be closely scrutinized." Guardians, 630 F.2d at 100 (citing to the Uniform Guidelines). The Court cannot find that the City has met this heightened validity standard. The City failed both to test a sufficient range of critical KSAs in the 2008 exam, and to produce evidence sufficient for the Court to conclude that the exam was valid.

Dr. Campion testified that the 2008 exam was sufficiently valid to use for rank ordering. He cited to the 2000 job analysis, in which SMEs, through the ranking process, concluded that the "more knowledge you have the better the lieutenant you could be." 12/19/14 Tr. 47:8-18. He also testified that thousands of studies called meta-analyses revealed, "based on a statistical relationship," that "higher scores [on knowledge tests] lead to higher job performance." Id. at 47:20-48:5. Both arguments are unpersuasive.

First, the fact that the 2008 exam measured only knowledge areas that differentiate among performers cannot compensate for the 2008 exam's failure to test for critical non-knowledge skills and abilities. It is possible that someone might excel on a portion of an exam testing knowledge, but tank on a portion

of an exam testing skills and abilities, and vice versa. If a test only examines knowledge (even if limited to knowledge areas that differentiate performance) while ignoring a broad swath of necessary skills and abilities, it hardly seems plausible that a higher score is likely to result in higher job performance, or even that the procedure measures aspects that differentiate among levels of job performance. What the Court can conclude from the 2008 exam is that those who excelled at the exam would exhibit superior levels of knowledge on the job, and that the 2008 exam differentiated among levels of candidates' knowledge levels. The Court agrees with Dr. Wiesen, however, that this is insufficient for predicting who will be a good police lieutenant. Wiesen Rebuttal 28.

Second, testimony as to the criterion validity of knowledge tests in general is insufficient to support the "refined use" of this exam as necessary for rank ordering. United States v. City of New York, 637 F. Supp. 2d 77, 131 (E.D.N.Y. 2009); see also Bradley, 443 F. Supp. 2d at 173 (noting that although cognitive ability is correlated with job performance, there was "no persuasive evidence in [the] record that the use of the written cognitive examination as the sole basis for rank ordering entry-level firefighter candidates [was] a valid selection procedure"). In other words, Dr. Campion's testimony on this

point was far too general to support the City's claim that the 2008 exam was valid enough to be used to rank candidates.

The Court also concludes that the City cannot use the 2008 exam to rank candidates because the City has failed to demonstrate that the 2008 exam was reliable. Reliability measures "the extent to which the exam would produce consistent results if applicants repeatedly took it or similar tests." Guardians, 630 F.2d at 101. This is especially important when a test is used to rank candidates. 12/16/14 Tr. 75-76; 12/19/14 Tr. 18. Courts do not require perfect reliability, but "[w]ithout some substantial demonstration of reliability it is wholly unwarranted to make hiring decisions, with a disparate racial impact, for thousands of applicants that turn on one-point distinctions among their passing grades." Guardians, 630 F.2d at 101. One way of demonstrating reliability is by showing how skillfully the questions on the exam have been formulated, which can be done by twice giving a sample of an exam to the same population. Id. at 102. Employers can also use split-half correlation, observing how consistently an individual scores on each half of the test. Id. Employers can do such analyses before or after a test is administered. Id.

The only evidence the City presented on reliability was HRD's determination in the 1991 report that the 1991 exam had a reliability rating of .79. 12/19/14 Tr. 42-44; 1991 Validation

Report at 00382. Dr. Campion claimed that the reliability of the 2008 exam was likely comparable to the 1991 exam because the exams were highly similar. Campion Report 15. The 1991 validation report is insufficient to demonstrate reliability. In addition to its age, it was evaluating a different test. The City's suggestion that the 2008 and 1991 exams were highly similar is inconsistent with the City's reliance on the 2000 job analysis during the trial, and with the 2005 and 2008 mini-job analyses. The City seems to suggest that despite all of these updates to the 1991 report, the exam hardly changed at all. The Court does not find this argument persuasive, and concludes that the City has not demonstrated the reliability of the 2008 exam. Further undercutting a finding of reliability is the fact that Dr. Jacobs (of EB Jacobs), recommended in 2009 that the City "band" the results of the 2008 exam in nine-point increments because "a given candidate's score will vary from one administration of a test to another." Lopez, Exs. 70-71. The Court cannot find that the 2008 exam was sufficiently reliable for use as a ranking mechanism.

4. Conclusions Regarding Job Relatedness and Consistency with Business Necessity (Prong 2)

The City has failed to demonstrate that "a higher score [on the 2008 exam] is likely to result in better job performance." 29 C.F.R. § 1607.14(C)(9). The state statute that the City

argues requires civil service employers to promote in strict rank order cannot serve as an affirmative defense for the City: Title VII relieves employers of state hiring requirements "which purport[] to require or permit" any discriminatory practice. Guardians, 630 F.2d at 104-05 (quoting 42 U.S.C. § 2000e-7); Bridgeport Guardians, Inc., 933 F.2d at 1148 (quoting 42 U.S.C. § 2000e-7); Vanguard Justice Soc., Inc. v. Hughes, 592 F. Supp. 245, 268 (D. Md. 1984) (citing Guardians, 630 F.2d at 104). Moreover, current Chief Judge Saris noted in Bradley that it was not clear that the statutory scheme prohibits banding.[41] 443 F. Supp. 2d at 174 (noting that "[w]hile the attorneys have not briefed the issue, banding based on scores that have no statistical difference to diminish the adverse impact of a rank-order system seems consistent with the statutory scheme and applicable caselaw under Title VII").

[41] The Commonwealth attempted to use band scoring in 2009. 01/06/15 Tr. 98:16-99:7. HRD declined to band because of a challenge by the Civil Service Commission. Id. at 99:18-100:2. In that same year, a Massachusetts Superior Court issued a preliminary injunction enjoining the Department from banding scores, pending HRD utilizing its rule-making authority to establish the banding process. See Mem. Decision Order Pls.'s Mot. Prelim. Inj. Order, Pratt v. Dietl, No. SUCV2009-1254 (Suffolk Sup. Ct. April 15, 2009), ECF No. 7. The Superior Court decision does not justify the City's rank ordering of candidates based on the 2008 exam; the ruling was a preliminary injunction, not a final judgment on the merits, and, in any event, it cannot override the requirements of Title VII.

This Court holds that even were the 2008 exam valid enough to be used as a screening tool,[42] the City has failed to meet its burden of showing that the 2008 exam was sufficiently valid to be used as a basis for ranking candidates. The City has

[42] Judge O'Toole held in Lopez that the 2005 and 2008 sergeant exams were "minimally valid" and the City had therefore met its burden of demonstrating that the 2005 and 2008 sergeant promotional exams were "job related" and "consistent with business necessity." Lopez, at 35-36 (internal quotation marks omitted). Considering the substantial overlap between the questions appearing on the 2008 sergeants' exam and those appearing on the 2008 lieutenants' exam (sixty-eight questions in common), Wiesen Rebuttal 8, a decent respect for the views of a colleague compels this Court to comment on the divergent holdings.

Judge O'Toole's findings are not so different from this Court's. Agreeing with the City's expert, Judge O'Toole held that, but for the E&E component, the 2005 and 2008 sergeants' exams would not have passed muster under Title VII. Lopez, at 35-36. This Court arrives at a different holding in large part because the Court does not place as much stock in the E&E component as did Judge O'Toole. First, as discussed above, the automatic award of fourteen of the possible twenty points on this component reduces its usefulness. Second, the City bears the burden of establishing the validity of the E&E component, meaning it must establish a relationship between the content of the training/experience that is rewarded, and the content of the job. 29 C.F.R. § 1607.14(C)(6). The Court could possibly infer such a relationship but, as the case law indicates, this is an area for which the Court looks to expertise. Dr. Wiesen points out, and this Court agrees, that the City has presented no evidence on this point. Wiesen Rebuttal 63. While SMEs did assist DPA in evaluating and developing the E&E component, there is no evidence linking the E&E inputs with tasks or KSAs from the job analyses. Lopez, Ex. 41 Apps. W and X. Lastly, E&E scores were calculated only for candidates who passed the exam. 12/15/14 Tr. 62:24-63:2. There was disparate impact in the pass rates for the 2008 exam. In light of Title VII's purpose, the Court is skeptical that a component of the test that was disproportionately unavailable to minority candidates can rescue an otherwise invalid exam.

therefore failed to meet its burden on the second prong of the legal framework: it has not convinced the Court that the 2008 exam was "job related for the position in question and consistent with business necessity." Jones, 752 F.3d at 54 (internal citation omitted).

The Court need not proceed to step three of the disparate impact analysis: the Plaintiffs have won their case.[43]

---

[43] The City's post-trial submission argued that language in a recent Supreme Court case changed the usual disparate inquiry, imposing a new rule: "a Court cannot reject a governmental defendant's Prong 2 justification if a plaintiff has not . . . presented enough evidence to meet its Prong 3 burden[.]" Def. City Boston's Supp. Post-trial Filing 5-6, ECF No. 196 (citing Texas Dep't of Hous. & Cmty. Affairs v. Inclusive Cmtys. Project, Inc., 135 S. Ct. 2507, 2518 (2015)). Admittedly, the Supreme Court's statement in dictum, read in isolation, does indeed appear to state that rule. See Texas Dep't of Hous. & Cmty. Affairs, 135 S. Ct. at 2518 ("[B]efore rejecting a business justification -- or, in the case of a governmental entity, an analogous public interest -- a court must determine that a plaintiff has shown that there is an available alternative practice that has less disparate impact and serves the entity's legitimate needs.") (internal citation omitted).

The Court notes that the Supreme Court majority favorably cited Title VII precedent and made no indication that it was changing a decades-old three-prong doctrine. See, e.g., id. (noting "cases interpreting Title VII . . . provide essential background" in construing other federal antidiscrimination statutes). More importantly, although it is far from obvious, the best reading of its opinion is that the Supreme Court was actually discussing the proper application of Prong 3 when it made that statement.

The Supreme Court affirmed the Fifth Circuit's reversal of the trial court. See id. at 2526. The Fifth Circuit had held that the trial court had conflated prongs 2 and 3 by placing the burden on the defendant to "prove that there are no other less discriminatory alternatives to advancing their proffered interests[.]" Id. at 2514 (internal citation omitted). Thus the Court's statement can be interpreted as merely reiterating

VI. CONCLUSION

Judgment shall enter in favor of the Plaintiffs regarding liability under Title VII and Chapter 151B for the 2008 lieutenant promotional exam. By order dated November 10, 2014, the Court bifurcated this trial into a liability phase and a damages phase. The first part is complete.

The Court and the parties must now proceed to the remedial phase of the lawsuit. The Court invites the parties to reach an agreement concerning an appropriate remedy. If they cannot, the Plaintiffs shall propose a remedy within thirty days of the date of this order, and the Defendants shall respond within thirty days thereafter.

---

that, if the defendant has met its burden on prong 2, the plaintiff still bears the burden on prong 3. Cf. Abril-Rivera v. Johnson, 795 F.3d 245, 253-54, 255-56 (1st Cir. 2015) (stating the familiar three-prong burden-shifting framework for disparate impact cases; discussing Texas Dep't of Hous. & Cmty. Affairs without revisiting the framework).

That is not to say the City's argument lacks force. In fact, it might actually ensure Title VII litigation helps counteract unjustifiable disparate impact, instead of merely negating it. Forcing Title VII plaintiffs to identify a workable alternative that would achieve more equal opportunity before invalidating an employer's practices under prong 2 would ensure that an employer's impermissible employment practice is replaced by one that is just as effective but which offers more equal opportunity. But under the current scheme, the plaintiff bears no burden of demonstrating the existence of equally effective alternative practices unless the defendant has successfully met the requirements of prong two.

SO ORDERED.

/s/ William G. Young
WILLIAM G. YOUNG
U.S. DISTRICT JUDGE