120

COMMONWEALTH OF MASSACHUSETTS

SUFFOLK, ss.

SUPERIOR COURT
C.A. NO. 0984CV00576

SPENCER TATUM, GWENDOLYN BROWN, LOUIS ROSARIO JR., and FRANCISCO BAEZ, individually and on behalf of a class of individuals similarly situated,
Plaintiffs,
v.

COMMONWEALTH OF MASSACHUSETTS, and PAUL DIETL, in his capacity as Personnel Administrator for the Commonwealth of Massachusetts, Human Resources Division,
Defendants.

FINDINGS OF FACT AND CONCLUSIONS OF LAW ON PHASE I (LIABILITY)

The best test-takers are not necessarily the best police sergeants. Yet, the Commonwealth of Massachusetts through the Personnel Administrator, Human Resources Division ("HRD") regularly administered written exams, knowing that its testing format had an unnecessary, plain and obvious adverse impact upon Blacks and Hispanics, compared to White candidates. To challenge HRD's format, a class of Black and Hispanic police officers (some now retired) filed this case in 2009, alleging racial and national origin discrimination in employment (G.L. c. 151B, § 4) in the police sergeant promotional examinations administered by HRD for the years 2005, 2006, 2007, 2008, 2010, and 2012.

The court conducted a bench trial in phase I of this case, limited to liability. It heard testimony from live witnesses and received over 300 exhibits on June 26, 27, 28 and 29 and on July 25, 26, 27 and 28, 2022. It heard arguments on July 29, 2022, and September 30, 2022. It received written post-trial briefs on September 15, 2022.

The evidence is very clear. It defeats any justification for HRD's heavy reliance upon biased exams to identify the best candidates for promotion to sergeant. Moreover, HRD knew of clearly superior assessment methods, but continued to use the same, unnecessarily discriminatory format anyway. The massive amount of evidence proving the known and unjustified disparate impact of HRD's format leaves no doubt in this court's mind that the Commonwealth has interfered with the plaintiffs' rights to consideration for promotion to police sergeant without bias due to race or national origin. G.L. c. 151B, § 4(4A). The court will therefore conduct phase II of this trial, in which it will determine a remedy. The remedy will provide relief to the plaintiff class, which must be commensurate with the deep-seated illegality in the testing format that HRD used, at least for the 2005, 2006, 2007, 2008, 2010, and 2012 exams.

In fashioning a remedy, the court realizes that candidates and appointing authorities relied upon and participated in HRD's process in good faith, although some agencies did seek a better way. The court in no way faults those candidates and appointing authorities. Least of all does the court cast doubt on the qualifications of the successful candidates. HRD's violation involved selection among qualified candidates, all of whom abided by HRD's rules. No party has suggested a remedy that would affect any existing appointments.

PRIOR PROCEEDINGS

This case travelled a long road to get to trial. In 2007, some of the plaintiffs sued the Commonwealth, HRD and their employing municipalities in the United States District Court for the District of Massachusetts. Lopez v. City of Lawrence, U.S. Dist. Ct. No. 07-11693, 2014 WL 12978866 (D. Mass. September 5, 2014), aff'd 823 F.3d 102 (1st Cir. 2016), cert. denied, 137 S.Ct. 1088 (2017) ("Lopez I"). In an interlocutory appeal, the First Circuit held that the state defendants were not "employers" within the meaning of Title VII and were therefore entitled to

immunity under the Eleventh Amendment to the United States Constitution." Lopez v. Massachusetts, 588 F.3d 69, 72 (1st Cir. 2009). The plaintiffs dismissed their state law claims against the state defendants under G.L. c. 151B without prejudice and refiled them in this court.

This court dismissed the entire case, relying in part on the 2009 First Circuit decision. The plaintiffs appealed. The Supreme Judicial Court affirmed dismissal of several claims, but rejected the defendants' sovereign immunity claims and held that the complaint stated a claim upon which relief could be granted on a theory of interference with protected rights under G.L. c. 151B, § 4(4A). Lopez v. Commonwealth, 463 Mass. 696, 701-702, 706-712 (2012) ("Lopez II"). On remand, on September 16, 2013, this court (Fabricant, J.) certified the class, finding that the case "presents questions of both law and fact that are indisputably common to all members of the proposed class, including the legal questions of what exactly plaintiffs must prove to show discriminatory impact and harm, and the factual questions of whether the test had a discriminatory impact in each of the years alleged."

Meanwhile, the federal case against the employers went to trial. The Federal District Court judge, sitting without a jury, found that, while the Boston promotional examinations caused a disparate impact based on race in 2005 and 2008, the tests were nevertheless job-related and consistent with business necessity, and the plaintiffs failed to prove that Boston had refused to adopt an alternative with less disparate impact. The First Circuit affirmed in Lopez I.

On June 27, 2018, Defendants filed their Motion for Judgment on the Pleadings, arguing issue preclusion based upon the First Circuit's decision in Lopez I. On January 7, 2019, the Superior Court (Tochka, J.) allowed the Motion and dismissed the Third Amended Complaint. The plaintiffs were again successful on appeal, this time from a judgment entered on January 14, 2019. In Tatum v. Commonwealth, 98 Mass. App. Ct. 1105, 2020 WL 4200865 at *2-*3 (2020)

(Rule 23 0 order), the Appeals Court held that the plaintiffs failed to establish identity of the parties, privity (because the federal district court denied class certification) and adjudication of identical issues. Upon remand, the case was specially assigned to the undersigned for trial.

FINDINGS OF FACT

The Plaintiffs are a class of current or former police officers who took a police sergeant promotional exam created and administered by the Commonwealth of Massachusetts Human Resources Division ("HRD") in 2005, 2006, 2007, 2008, 2010, or 2012. The Plaintiffs include current and former officers for the cities of Brockton, Lawrence, Methuen, Lowell, Springfield, Worcester, Boston, and other cities and towns throughout the Commonwealth of Massachusetts, as well as the Massachusetts Bay Transportation Authority ("MBTA"). Each named and class Plaintiff is either Black or Hispanic. The Plaintiffs were either not promoted to sergeant or experienced a significant delay in such promotion based on their scores on HRD's examinations.

The court finds the following facts by a preponderance of the evidence it finds credible:

I. The Tests and Promotional Lists

1. HRD developed and administered all of the examinations at issue in this action (the 2005, 2006, 2007, 2008, 2010, and 2012 exams). The exams were substantially the same in format and approach, though the specific questions were different.

2. The sergeant's promotional examination in 2005 through 2012 consisted of 80 multiple-choice questions. For the same years, HRD simply added 20 questions for the lieutenant's exam, and 20 more questions for the captains. The higher rank exams thus included all of the questions on the sergeant's exam.

3. The written examination's multiple-choice questions were all taken (sometimes verbatim) from police-related textbooks. This component has been in effect for at least 50 years.

4 The educational and experience ("E&E") component nominally accounts for 20% of a candidate's overall score. It is essentially the same today as it was 50 years ago.

5 The exams generated results that were largely reproducible ("reliable"). HRD's expert, Dr. Silva, calculated the reliability of the exams using "Cronbach's alpha", which took the average of all split halves of each respective exam. He found that the lowest reliability of any exam was .71 and the highest was .84. The court concludes that all of the exams had good reliability.

6 Based on the exams, HRD created lists of candidates, ranked by order of their scores, for use by appointing authority. Each appointing authority used the list to promote candidates from within the ranks of that municipality's police force. Police officers from each participating municipality for any of the exams at issue competed only against the other officers from their department for spots on the eligibility list for their municipality. For instance, under no circumstances can a police officer from Chelsea be promoted to sergeant in the Quincy police department.

7. Each of the examinations at issue in this case contained different questions, had different municipalities participating, and with the exception of repeat test takers, had different candidates taking the examinations. HRD issued different eligibility lists for each participating municipality for each of exams at issue (i.e., the 2005, 2006, 2007, 2008, 2010, and 2012 statewide exams, and the 2005 and 2008 Boston exams).

Test Development

8. The outlines used for the 2005, 2006, 2007, and 2008 statewide sergeant's exams were based on the knowledge, skills and abilities ("KSAs") and job tasks identified as important to the job of sergeant in the Validation Report for the 1991 Police Promotional Selections Procedures dated October 1, 1991 ("1991 Validation Report") and the 2000 Morris & McDaniel Job Analysis Report. There were some major shortcomings in HRD's use of these Reports, as discussed below.

9 The outlines used for the 2005, 2006, 2007, and 2008 statewide exams included KSAs in the following categories for sergeant: Law/MGL, Supervision, Community Policing, and Police Functions.

10. Guy Paris, who began working in test development for HRD around 1990, was primarily responsible for writing test items, or multiple-choice questions, for the 2005, 2006, 2007, and 2008 statewide exams.

11. To write questions for those statewide exams, Mr. Paris consulted an outline that identified the competency areas by category that were important for the job of sergeant. The outline listed the number of questions to be written for each competency, and incorporated and linked these questions to the source material on the reading list for each respective exam.

12. Mr. Paris made sure that the reading lists for the 2005, 2006, 2007, and 2008 statewide exams listed the most current versions of the sources as of the reading list publication date.

13. In February, 2005, prior to administering of the 2006 exam, HRD surveyed 170 community police chiefs about the proposed reading list. The police chiefs rated the use of particular sources and recommended other sources.

14. There were 80 multiple-choice questions on the 2005, 2006, 2007, and 2008 statewide exams, but Paris wrote many more than 80 questions so that subject matter experts ("SMEs") could review the potential questions and select the most appropriate questions. The use of SMEs is a best practice in developing the written portion of a police promotional exam.

15. For those statewide exams, HRD retained two to four SMEs, who were typically police chiefs, to review the potential questions. For instance, Robert Champagne, the former Police Chief for the City of Peabody, served as one of the SMEs from 2005 to 2012. HRD rarely used more than three SMEs and, in most cases, appears to have used only two SMEs, both of whom were police chiefs, to review reading lists and examination questions. With so few SMEs and given the deficiencies identified below, the court gives only modest weight to the SME process in assessing the validity of the exams for statewide application or use in Boston.

16. After Mr. Paris drafted questions for the 2005, 2006, 2007, and 2008 statewide exams, he conducted meetings in which the SMEs reviewed the potential questions and rated them for suitability to each rank, difficulty, readability, and recommended use. The SMEs also reviewed them for content, consistency, applicability, and practicality.

17. An HRD consultant, E.B. Jacobs also reviewed the potential questions for the 2007 and 2008 statewide exams for cultural bias. It revised some questions and recommended replacing some questions.

18. In the summer of 2009, as part of a mini job analysis for the police promotional exam, which included the sergeant's exam, HRD's manager responsible for the 2010 and 2012 exams, Lauren Fitzgibbons, met with SMEs to review the reading list and the exam outline.

19. Those SMEs recommended that the criminal law and constitutional law guides from Attorney Rogers of Commonwealth Police Services, Inc. be used in place of the criminal

law and constitutional law guides that had been listed on the reading lists in previous years. Ms. Fitzgibbons also provided the SMEs' comments about the reading list and the outline to E.B. Jacobs for use in writing the questions for the 2009 statewide exam.

20. In 2009, Ms. Fitzgibbons also held meetings with the SMEs to review proposed questions for use on the 2009 statewide exam.

21. Based on SME input, Ms. Fitzgibbons revised the reading lists for the 2010 and 2012 statewide exams to include Massachusetts Criminal Law and Massachusetts Criminal Procedure by Attorney Rogers.

22. E.B. Jacobs, created the outlines for the 2010 and 2012 statewide exams, which included KSAs in the following categories for sergeant: Law/MGL, Police Functions, Community Policing, and Supervision.

23. Through these and other steps, HRD kept the reading lists and resulting exams current. Within the constraints of a written multiple choice test and limited E&E component, HRD's exam development process was comprehensive and conscientious.

24. The court rejects the inference that these test development processes, by themselves, "ensured that those exams were job related" (Def. Prop. Findings VII). Even apart from the inability of a written multiple-choice exam to predict good job performance as a sergeant, a good process may be a necessary component of job-related testing, but it is not sufficient, as even a brief summary will show. For instance,

a) The questions on the exam largely test for rote memorization of facts and passages taken directly from textbooks that candidates are asked to study. The 1991 Validation Report and 2000 study did not identify test-taking skills and lack of test-related anxiety as job-related.

b) HRD also had candidates complete a computerized form E&E worksheet, which assigned candidates certain points based on certain educational criteria (discussed fully below).

c) HRD gave 80% weight to a candidate's scores on the multiple-choice exam and 20% weight to the E&E component. No credible study validated these weights.

d) For each exam, HRD set a passing score for the multiple-choice test. In all but one of the challenged exams, HRD set the passing score at 70. It did not rely on any accepted scientific criteria for establishing the passing score for its exams.

e) Once HRD tabulated exam scores, it created a promotional eligibility list for each department that participated in the exam. It did so in rank order, according to scores rounded to the nearest whole number. No credible study showed that single-point differences in scores reflected any significant difference in job qualifications.

25. At the time it administered the exams in question – and for a long time before that – HRD knew that police departments throughout the Commonwealth generally promote candidates on the eligibility list in rank order fashion.

26. A one-point difference in exam score can make the difference between promotion and being passed over. It can also cause denial or delay of promotion. It can also make the difference between being considered for promotion and excluded from consideration. HRD knew this when it administered the tests at issue in this case.

27. The court now turns to a more detailed discussion of HRD's testing and ranking format, roughly following the same order that HRD used in administering that format.

Knowledge, Skills, and Abilities ("KSAs")

28. A valid exam targets and measures the important KSAs needed effectively to perform the position at issue, and then assesses at least a representative sample of the most important skills.

29. Police sergeants are first-level supervisors, with direct responsibility for supervising patrol officers on a day-to-day basis. They are responsible for responding as needed to routine calls and must respond to all serious incidents (e.g., aggravated assaults, homicides, shootings, sexual assaults, community disorders involving racially-motivated incidents, armed robberies, and injured officers).

30. Chief among the essential skills and abilities for this position are:

- Leadership skills;
- Supervision skills;
- Decision-making and problem-solving;
- Interpersonal skills;
- Communication skills; and
- Integrity.

31. A number of studies and reports confirm the primacy of these KSAs. For instance HRD's 1991 Validation Report determined that critical abilities of a police sergeant include, among other things, (a) "ability to make and carry out decisions quickly," (b) "ability to give clear, concise verbal orders," (c) "ability to communicate orally and in writing," (d) "ability to bring calm to control surroundings when in stress producing situations," and (e) "ability to establish rapport with persons from different ethnic, cultural and/or economic backgrounds."

32. The court also accepts and adopts as fact the statements of a subject matter expert, former Peabody Police Chief Robert Champagne. He stated accurately that good sergeants "could communicate well, people that had – that were approachable, people that had a nice demeanor, people that commanded respect from the people that were there that were

knowledgeable that the 1991 Validation Report knew what it was that they were talking about." The court also adopts as fact Chief Champagne's testimony that knowledge of the profession and "just general knowledge of the – of the city" were desirable traits for aspiring sergeants.

33. The 1991 Validation report generated a comprehensive list of KSAs required to perform the job of police sergeant. It included a job analysis study that included the following major steps: (1) "[g]athering of available job information from Massachusetts police departments, as well as job analysis reports, survey instruments, and other information from jurisdictions outside the Commonwealth"; (2) "[d]evelopment and administration of a task inventory questionnaire designed to identify the frequent and critical tasks and duties of each of the five ranks," including sergeant; (3) "[d]evelopment and administration of a knowledges, skills, abilities and personal characteristics (KSAPs) inventory questionnaire designed to identify the important KSAPs required at the time of appointment to each of the five ranks," including sergeant; (4) "[l]inkage of the important KSAPs to the frequent and critical tasks of these jobs by subject matter experts (SMEs)"; (5) "[d]esign and use of critical incident technique (CIT) structured group discussions to gather from SMEs descriptions of actual incidents which have occurred on the job"; (6) "[d]esign and use of structured group discussions to gather information from SMEs about the Education and Experience (E&E) component of DPA's selection procedures"; and (7) "[d]esign and use of structured group discussions to gather information from SMEs about the recommended reading list, from which the multiple-choice written examination questions for the police promotional exams are derived."

34. From the inventory of 187 total KSAs, the SMEs identified 159 KSAs in the 1991 Validation Report that were needed to perform the job of sergeant. The 1991 Validation Report

was substantially accurate and reliable in identifying the necessary KSAs for the police sergeant position.

35. The 1991 Validation Report, however, has serious flaws in identifying (1) which KSAs are testable on a multiple-choice exam and (2) which KSAs are measured in HRD's education and experience component.

36. A 2000 job analysis conducted by the testing firm Morris & McDaniel for the Boston Police Department determined that critical KSAs of a police sergeant include, among other things, (a) "skill in supervision and leadership," (b) "skill in implementing community policing procedures and techniques," (c) "interpersonal skills," (d) "presentation skills," (e) "oral communication skills," (f) "ability to remain calm in stressful situations," (g) "ability to instill confidence," and (h) "ability to think under pressure." The court adopts these determinations as to Boston.

37. The 2000 job analysis assessed only the KSAs required of a police sergeant in the Boston Police Department, and did not analyze the KSAs of police sergeants in other municipalities or departments.

38. On rating sheets, the SMEs ranked the importance of various KSAs as part of that job analysis. Their rankings are implausible. The 11 SMEs gave identical rankings to all of the approximately 1,100 ratings, which is all but impossible in the absence of coordination between the SMEs. The court does not credit those rankings, which almost certainly reflect some unknown factor that interfered with the SMEs' independence.

39. HRD made no attempt to defend the inexplicable unanimity of these rankings, aside from speculation. There is simply no evidence demonstrating that any complete consensus

occurred and given that each of the questions instructed the rater to answer based on the work that they do, it is not possible for everyone to have agreed unanimously on every ranking.

40. The 2000 job analysis also claimed that police sergeants perform certain tasks every day, but that could not possibly be true.[1] Those tasks included:

- Qualifies and/or engages in required practice of operation of firearms and other weapons;
- Investigates and resolves citizen complaints against police officers;
- Set up command post at scenes of robberies, homicides, fires, etc.;
- Directs activities at the scene of major incidents (e.g., serious/fatal accident, crime, natural disaster, etc.)
- Conducts internal investigations;
- Investigates and prepares reports regarding misconduct by subordinates;
- Investigates use of force and injury to prisoner incidents and prepares reports for superiors as required;
- Recommends subordinates for commendations and disciplines them for dereliction of duty;
- Inspects licensed premises and prepares reports on violations, if any are found; and
- Talks with leaders of demonstrations.

41. The 2000 job analysis did reach some highly significant conclusions. Most importantly, it determined that at least half of the skills necessary for the job of police sergeant should be tested by assessment mechanisms other than a written multiple-choice examination. The court adopts that determination.

42. Finally, HRD never performed a criterion validity study, which is designed to evaluate the extent to which a sergeant's actual job performance correlates with performance on an examination. A criterion validity study would have been feasible where HRD has used the same examination format for decades and there is thus a large pool of current and former

[1] Since the job analysis asked 11 sergeants with diverse job assignments to state what tasks they themselves did every day, these claims would not be true even if, for instance, some sergeants in a specialized unit (e.g. internal affairs) did conduct internal investigations daily.

sergeants within the Commonwealth whose job performance and test scores could have been analyzed to evaluate the validity of HRD's promotional exams.

Written Exams

43. Most of the questions on the exams at issue in this case tested topics that were important to the job of sergeant. That does not mean that HRD's format was reasonably job related. It was not.

44. Because HRD failed to test many important KSAs, measured test-taking skills and memorization, enabled test-related anxiety to affect results and failed to ask questions that focused upon measuring job-related knowledge, its format did not rank candidates for promotional purposes on a basis that was substantially job related.

a. Testability of KSAs

45. For one thing, the exams did not test many important job qualifications. More importantly, written questions on a particular topic often test details, rather than job-related knowledge of information and principles actually used on the job. Testing for knowledge of soon-forgotten details does not measure ability to apply knowledge practically and to exercise judgment on that topic in specific situations, as a sergeant actually does on the job. Finally, the tests measure a candidate's test-taking skills, abstract knowledge and ability to memorize source material. A sergeant does not need these skills in practice. Nor does a candidate need abstract knowledge that does not reflect the ability to use judgment in a practical way on the job.

46. As part of the 1991 Validation Report, the SMEs identified 58 KSAs that, in their view, were tested by the multiple-choice component. HRD did not develop alternative methods of evaluating the remaining KSAs. It simply opted not to test them at all. Yet, many of the critical skills and abilities could not possibly be tested in a written multiple-choice examination.

These included "[s]kills in identifying problems, securing relevant information from both oral and written sources, identifying possible causes of problems, and analyzing and interpreting data and complex situation [sic] involving conflicting demands, needs, or priorities," "[a]bility to confront problems, take charge, and assume responsibility," "ability to appropriately delegate assignments," "ability to plan," and "ability to develop alternative solutions to problems and evaluate courses of action and to reach logical decisions based on the information at hand," among others. Those skills, among others (such as ability to communicate orally with subordinates and civilians), call for situational judgment and interpersonal skills, rather than theoretical and academic knowledge about such judgment and skills.

47. Implausibly, HRD stated the KSAs just quoted could be tested adequately and appropriately on the written examination, with questions in the format HRD actually used. That was not true.

48. Moreover, as written, HRD's actual examinations did not in fact test even for some skills that could have been tested, because some questions addressed abstract knowledge and failed to focus on matters relevant to performance as a police sergeant.

49. Where HRD did address questions of empathy or the dangers of authoritarian supervision, it did so by asking an informational question, rather than testing whether the individual candidate had empathy or authoritarian traits. Thus, reaching the correct answer turned upon test-taking skills, temporary memorization, or academic understanding of facts unrelated to actual job performance. For instance, question 35 on the 2010 exam asked:

> According to CP, the most critical determinant of future success as a community policing Officer is:
> A. superior communication skills.
> B. Empathy.
> C. autonomy.
> D. Analytical ability.

There is no reason to think that a candidate who knows that the correct answer is "B" will actually have more empathy than someone who thinks that a plausible alternative answer is what "CP" lists as the "most critical determinant."

50. Likewise, question 37 on the same exam asked:

> Barker and Carter found that authoritarianism is a dominant trait among Officers. According to CP, police managers should:
> 1 Recognize that authoritarian traits are most prominent in young Officers and that they tend to subside with experience.
> 2 Attempt to reduce authoritarianism and its behavioral consequences because Officers tend to become more authoritarian over time.
> 3 Encourage Officers to take an authoritarian approach because it often helps to control a situation.
> 4 Not involve an Officer possessing this trait in community policing efforts because it will likely escalate the Officer's degree of authoritarianism.

When asked about this question, the Commonwealth's SME responded that authoritarianism was a problem, but not between 2006-2012 when, in his opinion it was getting better. He pointed out that, most important, a sergeant should identify whether authoritarianism was a problem and, if so, to train the officer(s) in question. The question did not address that most important skill and, according to the subject matter expert, may be addressing a largely outdated concern.

51. HRD long knew that many important KSAs could not be tested in either the written or E&E component. When it oversaw the Morris and McDaniel's job analysis in support of the 1987 Boston Police Department promotional examination and set the allocation of points across various components of the examination, it knew that the experts at that firm believed that the written test did not assess many of the attributes needed for the job and should account for no more than 40% of the overall score.

52. While the true percentage of KSAs that are testable through a written multiple choice test is open to some debate, the court accepts the range from (a) Morris and McDaniel's

1989 estimate that 40% of the KSAs could be tested in a written test to (b) Dr. Wiesen's estimate that HRD only tested 22% of KSAs in the multiple choice test.

53. In 2000, the Boston Police Department commissioned another job analysis by Morris & McDaniel, which led to the 2002 Boston examination. Morris & McDaniel advised HRD and the Boston Police Department to administer an examination that included non-written components (an assessment center and performance review system) that, collectively, received as much weight as the written multiple-choice examination:

> The method of evaluation of a candidate for promotion on a KSA may include, but is not limited to, a written examination, an assessment center, a training program, a probationary period, and/or a medical/physical examination. **The method of evaluation is dependent on the appropriateness of measurement for the particular KSA.** For example, knowledge of search and seizure laws can be evaluated most effectively in a written examination, whereas **ability to communicate orally is more appropriately evaluated through a performance based assessment technique such as an oral board or an assessment center.** [Emphasis supplied].

Ex. 42 at p. 1 (Bates number 1550). This advice is sound.

54. That advice is also consistent with the testimony, in *Lopez I*, of Ed Davis, the former Commissioner of the Boston Police Department. He testified that a written exam should be a component of every sergeant's exam because "the basic fundamental knowledge that's needed to be in a supervisory rank in a police department. . . is so important to the day-in-day-out work of [a sergeant]." .

55. Many of the KSAs identified in the 1991 job analysis and 2000 Morris & McDaniel study call for evaluation through "a performance based assessment technique." Failure to do so injects extraneous influences (such as test-taking ability and temporary memorization skills) into the selection process, while diminishing the exam's ability to measure important KSAs accurately or appropriately.

56. HRD's knowledge of the Morris & McDaniel 2000 study also confirms its knowledge that a written multiple-choice test alone does not sufficiently test for the skills and abilities necessary for the job.

57. The limitations of HRD's format in testing KSAs are also apparent when compared to alternatives. In 2002, HRD approved the City of Boston's plan to introduce a performance review system to the examination process. Under that system, candidates' prior job performance would be reviewed and assessed as part of the promotion process.

58. As Dr. Silva acknowledged at trial, performance review systems "can be useful and they do tend to reduce adverse impact." His own company, E.B. Jacobs has recommended use of such systems. However, Boston's plan to implement a performance review system was ultimately abandoned following opposition from the police unions. In scrapping the plan, then-Police Commissioner Paul Evans stated:

> Just as we have changed the way we do police work, we must change the way we promote. We need to understand that our promotional system remains mired in a tradition that has become obsolete and disconnected from the way we do business today. We must be willing to reward police work, not memorization skills.

b. Formulating Questions for Testable KSAs

59. The exams at issue also included questions that lack "fidelity," i.e. a relationship to a sergeant's job, even though the questions nominally relate to, for instance criminal procedure and criminal law. A patrol supervisor sergeant would use the criminal procedure and the criminal law portion of what HRD assigned candidates to read and study. The other assigned reading materials may cover important topics, but knowing those sources has considerably less relationship to the sergeant position. Apart from the criminal procedure and criminal law, the technical knowledge part of HRD's exams had only an attenuated connection, if any, to the actual job.

60. For instance, sergeants must apply their knowledge by exercising judgment in specific situations. It is possible to write situational judgment questions, and some examinations do so, but very few of HRD's questions do so.

61. HRD's multiple choice questions regarding topics covered in assigned source materials follow a common format: They start with "according to" the source, followed by approximately 4 sentences, followed by four choices, based upon the sentence. These questions test knowledge of the source material. Such an emphasis on memorization of source material lacks support in any analysis of KSAs needed to perform well as a police sergeant.

62. Similarly, many questions are definitional in that the answers turn upon the meaning of a particular word. Those questions have low fidelity, because a sergeant's job does not generally involve using academic jargon or other definitions of concepts in the assigned reading. Dr. Wiesen's estimate that 20% of the questions are primarily definitional is reasonable.

63. Some examples discussed during trial[2] illustrate and prove that testing for knowledge of material assigned and memorized during test preparation is not the same as testing for practical knowledge used on the job.

64. On cross-examination of plaintiffs' expert, the Commonwealth asked about questions 4, 5, 10 and 15 on the 2005 exam, which ostensibly measure knowledge of a relevant topic of criminal procedure. They do measure that knowledge, but not in a way adapted to

[2] The court recognizes, of course, that taking isolated examples from the voluminous record does not prove a trend or overall conclusion. The court therefore concentrates on examples chosen by the Commonwealth to justify its position during the trial, because those examples illustrate problems in the Commonwealth's justifications for the exams.

distinguishing between candidates who will have the substantive knowledge they need to be good sergeants from those who will not. Illustrative was question 15, which reads:

> 15. According to PA, the aim of the decision made by the Supreme Court in the case of Mapp v. Ohio was to:
>
> A. ban the use of illegally-seized evidence in criminal cases.
> B. affirm equal protection, including legal counsel, for all requiring it.
> C. handcuff police in their struggle with lawlessness.
> D. require police officers to inform suspects of their constitutional rights during the course of questioning.

Lopez Ex. 48, p. 8. Since answers A, B and D are arguably correct in many circumstances,[3] selecting the right answer requires knowing that the case announcing the exclusionary rule was called Mapp. v. Ohio. A sergeant does not need to know case names. Asking the question in the matter set forth in 2005 Exam question 15 may produce a "spread" among candidates, which is desirable from a question-writer's perspective, but that spread distinguishes between candidates on the basis of knowledge of names of decided cases, not on the basis of knowing that illegally-seized evidence is generally inadmissible.[4]

---

[3] One can argue that options B and D are worded too broadly to be correct, but then, some illegally seized evidence may be admissible in criminal cases in certain circumstances too. See, e.g. Grasso and McEvoy, Suppression Matters under Massachusetts Law, §§ 20-3[a], [b], [c], [d] (LexisNexis 2020) (discussing exceptions to the fruit of the poisonous tree doctrine: independent source, attenuation, inevitable discovery). While articulation of those doctrinal exceptions post-dated Mapp, it is certainly not important for a police sergeant to know the historical development of exceptions to the exclusionary rule; the sergeant needs to know only what rules apply to the time when the events are occurring.

[4] Other cross-examination about questions on the 2005 exam (#4, 5, 26, 27, 28) failed to establish that the questions distinguished between candidates based upon predicted job performance. Those questions were most likely to separate candidates based upon the degree to which they were "test-wise" and upon their ability to decipher convoluted questions. While question 4 does set forth a fact situation and asks for options, the court agrees with Dr. Wiesen that the question is not an effective situational question because it requires the candidate to state (rather academically and "according to MGL and MVL") what the "STRONGEST legal action that [the officer] can take" (however, unwise) as opposed to what the police "should" do. It is important to know what actions are unlawful, but not what excessive actions the officer could take without breaking the law. The question does not test judgment about the most appropriate response.

65. HRD conducts a post-exam review of questions that appeared to be problematic for applicants. This can result in disregarding a question altogether in the scoring or in deeming more than one answer correct. On the 2007 statewide examination, HRD determined that 18 of the 80 questions (or approximately 22%) were flawed and either eliminated or double-keyed them. This high proportion of flawed questions reflects deficiencies in exam design and question-writing.

66. The Commonwealth's subject matter expert's testimony was also revealing regarding question 19 on the 2008 exam, which reads:

> According to PDRCIP, regarding the requirements for a search to be legally valid, it would be correct to state that:
>
> A. Neither the search warrant nor the accompanying affidavit are required to be brought to the scene.
> B The affidavit must be brought to the scene and presented on demand, to the owner or occupant of the premises, but it is not required to bring a copy of the search warrant.
> C The search warrant must be brought to the scene and presented on demand, to the owner or occupant of the premises, but it is not required to bring a copy of the affidavit.
> D. Both the search warrant and the affidavit must be brough to the scene but the police are not required to present either document to the owner or occupant of the premises, even upon demand.

Chief Champagne explained why he found this question job-related, but cited practical, not legal reasons:

> Q Can you explain why you believe that's a really good question?
> A So, oftentimes, if you show up at somebody's house to serve a search warrant, they're going to say, I want to see the search warrant -- to make sure that, by the way, that we're not bluffing as the cops; that we actually have one, right? No. 2, to make sure we have the right house, because from time to time we make mistakes. We're supposed to be at 22; we're at 24, right, and so somebody says, that's not me. I'm not Mr. Johnson; I'm Mr. Jones. Johnson lives next door. So I think there's practical things that are right there for both the police and for the -- for it to be there, but I also think [9-168] that -- I think that the intent -- and, again, it's my opinion -- but I think the intent of the law was to do it the right way, right? Somebody wants to see that the Court has seen fit to give us a warrant to search their property, show it to them, why not? Q. Is that information that a sergeant would need to know? A. Abs -- I think it's -- I think that's -- yes, my opinion is

absolutely he'd need to know that for the sake of why escalate something? Why -- put something to rest.

Tr. 9-168. Chief Champagne agreed that it would "make more sense to ask the sergeant candidate what's good practice to bring to [the warrant to the scene] as opposed to what's legal, because it might be that good practice is more than what's required by the law." Tr. 9-168-169.[5]

67. Another example of information that fell within the general topic of criminal law, but was unrelated to the sergeant's job, was inquiry into the maximum length of prison sentences allowed by law for certain offenses (2008 Military Make-up Examination, questions 6, 7, and 8). Judges and criminal trial lawyers need to know this, but sergeants could look up this information in the unlikely event they needed to know.

68. These examples demonstrate the difference between the practical information that sergeants need to know and use in practice and the more academic information sought in the questions, such as case names, legal doctrine, or legal source for mandatory police practices. A single point difference in exam score may make the difference between being considered at all and being promoted - - and in being delayed in promotion. It therefore does not take many ill-conceived questions to make a difference.

69. Because the above-cited questions (and others) received SME approval despite the presence of significant demands upon applicants to apply non-job-related skills, the court infers that HRD did not adequately instruct its subject matter expert, or question writers to avoid focus upon memorization of abstract or academic information, or to determine and test the

[5] Indeed, question 19 may even be counterproductive, because there clearly are situations where having the warrant affidavit, though not required, has the practical benefit of clarifying the location to be searched in some unanticipated situations. See Commonwealth v. Hamilton, 83 Mass. App. Ct. 406, 414-415 (2013) (warrant affidavit clarified which apartment was to be searched); Commonwealth v. Toledo, 66 Mass. App. Ct. 688, 692-700 (2006) (warrant affidavit made clear that the warrant used the wrong street name).

information that a sergeant actually needs. HRD failed to instruct them that, if the subject matter expert relies upon practicalities rather than legal precedent, the question should do the same, because that is what real sergeants do.

70. It follows that, even for testable KSAs, HRD's actual questions test for extraneous skills, including academic-level understanding, attribution to specific source material, and test-taking skills. A candidate with equal knowledge of relevant, testable content may fare worse than an equally-qualified (or even less-qualified) candidate versed in test-taking, academics or temporary memorialization.

c. Precision

71. Even where HRD's questions actually tested for KSAs in the manner required to perform well as a police sergeant, the questions remain: to what extent, if any, does the final scoring warrant the conclusion that a candidate with a higher score is likely to perform better as a police sergeant than one who scores lower? Does a single point difference in scores meaningfully predict who is the better candidate? If not, does some larger point difference do so?

72. HRD's Eligibility Lists distinguish between candidates based upon scores, differing by one point or more, on a process almost entirely driven by scores on a rote-memory multiple choice test.

73. Given the deficiencies in the test, the point score differences are not job related, except perhaps where the differences are very large. There is no reason to believe that a candidate who scores one point higher than another candidate (or even 3 or more points higher) is likely to be a better police sergeant than the lower-scoring applicant.

74. HRD itself recognized that single-point differences in scores did not predict candidate qualifications and, in fact, proposed "banding." Banding entails grouping candidates falling within a range of scores and treating them as though they all received the same score. It reflects the reality (as the court finds) that a one-point difference in scores is not job-related.

75. As stated earlier, in a number of years (1992, 1996), it was in fact HRD and local departments, including the City of Boston and the MBTA, who took the position formally and in court that the promotional examination was not sufficiently valid to justify strict rank order selection, and that it was thus appropriate to hire out-of-order by selecting lower scoring minorities. In essence, HRD has already conceded that its multiple-choice examinations were not sufficiently valid as rank order devices, even though they now claim just the opposite.

d. Development of New Tests

76. For each test, HRD wrote new test questions every year and submitted those questions to SMEs for review and revision. The consistent references to written source materials did not change materially.

77. For the tests at issue, HRD did not vary the basic format of a rote-memory multiple choice exam and E&E with defined components. Nor did HRD change the underlying KSAs it identified. It therefore failed to test meaningfully the KSAs required for good performance as a police sergeant.

78. HRD assessed the knowledge, skills, and abilities ("KSAs") of police sergeants on a statewide basis, without differentiating between departments or municipality size or demographics, as part of its 1991 Validation Report.

79. The changes from exam to exam did not address the shortcomings in job relatedness identified above.

e. Police Sergeant Performance on Later Written Exams

80. Because Police sergeants do not use many of the KSAs tested by HRD's rote-memory multiple choice tests, incumbent sergeants who take HRD promotional exams for lieutenant often do not perform well on the sergeants' portion. Incumbent sergeants forget the abstract knowledge they acquired to pass the sergeant promotional test. The information that successful candidates memorized for the HRD promotional exam was not something they used regularly on duty as sergeants and therefore they did not commit it to memory.

81. This also confirms that memorization of the assigned reading material is not necessarily important to good performance as a police sergeant. Police sergeants can consult source information (or other individuals) to avoid mistakes, which may be advisable except in an emergency. While they may need to memorize some information, memory is unreliable. There is a substantial risk of mistakes on the job due to failure of memory if sergeants rely heavily on memorization. HRD knew all this at the relevant time. There is no credible evidence that it evaluated which information police sergeants must memorize in order to perform their job.

f. Influence of Extraneous Factors

82. Rote memory multiple choice tests, with questions drawn from a reading list, favor those whose educational experiences included such examinations. Familiarity with such tests likely reduce the anxiety of test-takers. Experience helps develop strategies for answering questions, including ways to identify "distractors" (false options) and identifying the methods used by question writers. HRD has not accounted for the existence of disparate educational opportunities and differing exposure to high-stakes rote memory tests as between racial and ethnic groups. The court credits the testimony of the witnesses in this case who pointed to

educational disparities as an explanation for the differing performance of such groups on HRD's tests.

83. Test taking skills are built through practice. Expert witnesses in this case observed that minorities, in general, have had fewer opportunities to participate in our educational system. This results in fewer opportunities for minorities to take tests and to become good test takers, which translates into the adverse impact seen on tests in general, especially tests of cognitive ability. Dr. Silva testified that he believes the difference in average performances is due to socioeconomic differences, lack of access to opportunity, and structural racism that exists within the system, all of which makes everything more difficult for minorities and impacts all tests. The court accepts and adopts this testimony.

84. HRD's format distinguishes in significant part between candidates based upon the educational and testing opportunities that the candidates had in the past, regardless of the candidate's personal strengths on the most important qualifications for performance as a sergeant, including the five most important skills and abilities identified above.

85. Given these potential explanations, scores on tests that differ by as little as one point lack a connection to potential ability as a police sergeant and may reflect fewer opportunities to acquire test-taking skills and to practice rote, temporary memorization for test-taking purposes.

Education and Training

86. As part of the 1991 Validation Report, SMEs reviewed the E&E component under which "incumbents receive points that in combination with their raw score on the written civil service exam become the score by which they are ranked and placed on civil service eligible lists of candidates for appointment."

87. The 1991 Validation Report did list some skills as being tested on the E&E component. Those skills included "perceiving and reacting to the needs of others", "ability to write, prepare reports", "ability to be confidential", "ability to follow policies and procedures", and "ability to interpret policy." Given the limited scope of the E&E component and the failure to consider or give credit for many elements of education and experience, HRD did not in fact capture these skills in the E&E component.

88. HRD considered a list of experience and education, limited to the matters and parameters set forth on an E&E Rating Sheet. Section III of the E&E Rating Sheet requested information on their candidate's work experience. Section IV asked candidates to report certain educational degrees. Section V of the E&E Rating Sheet asked candidates to list any courses above the high school level that they took in the same subject areas for which credit was given if an educational degree was earned in that area.

89. Sections III, IV and V did not assess how the candidate performed during earlier study and training in prior jobs what they learned or what skills they exhibited.

90. Neither Sections III, IV nor V nor any other aspect of HRD's format considered supervisory experience outside of law enforcement (such as private industry or the military). HRD did not grant credit for certain kinds of law enforcement training (e.g. in the military police or other military position). It gave no credit for matters such as community policing or involvement in the communities served by the candidate's department. It did not assess other experience that would provide useful background for a police sergeant.[6]

[6] Similar limits are in place for certain points awarded because of statutory preferences, see Ralph v. Civil Service Commission, 100 Mass. App. Ct. 199, 199 (2021) (no points for experience as an auxiliary police officer and as a special police officer).

91. HRD did not validate its treatment of education. For instance, there is no credible support for the notion that a bachelor's degree was the equivalent of six years job experience. Granting an incentive for officers to receive education (as in the Quinn bill) is not the same as validating qualifications for rank-ordering of candidates on a promotional exam.

92. As part of the 1991 Validation Report, SMEs identified 23 KSAs that, in their view, were tested by the E&E component. Given the limited scope of information considered on the E&E weighting, however, HRD's format did not actually assess many of those KSAs.

93. "Based upon an analysis of past standard practices and the most recent [E&E] weighting schemes used in [pre-1991] Boston Police Department promotional examinations which were developed by [SMEs], the decision was made to rate the [E&E] component at 20% of the final mark with the balance of the mark attributed to the written examination." 1991 Validation Report at 15. No empirical support or credible professional study justified the 20% weighting.

94. In practice, the effective weight of E&E component is substantially lower than 20% because of the way HRD scores E&E.

95. A candidate receives 14 of the 20 points just for being eligible to sit for the exam by virtue of having been an officer for three years. HRD has not cited any study linking the allotment of 14 parts to the relative importance of three years of seniority. A candidate receives the same 14 points whether he or she has three years or eighteen years of experience as a police officer.

96. The effective weight of the E&E component, therefore, is 6 to 8 points, not 20 points out of 100. In practice, a candidate's score is based nearly entirely on the written multiple-choice component, with the "education and experience" component having a minimal

impact on a candidate's score. HRD has long been aware that it has never validated its E&E rating methodology in accordance with any principles under the Equal Employment Opportunity Commission's Uniform Guidelines on Employee Selection Procedures (1978) (the "Uniform Guidelines").

97. In fact, the final scores on HRD's exams correlated in a perfect linear relationship with the score on the multiple choice tests, with statistically significant certainty.

98. In sum, by 2005 HRD knew that its written 80 question multiple-choice rote memory exam, when used as a rank order device, and without any other type of assessment center, even when coupled with the E&E rating, was not valid either under the Uniform Guidelines or M.G.L. c. 31 § 16.

II. Exam Scoring and Adverse Impact

Statistical Methods and Measures.

A. Statistical Significance and Adverse Impact Ratio

99. Statistical significance is a term of art. In this case, the experts assessed statistical significance by calculating a probability (a "p-value") for comparisons between candidates' performance and promotions. A "p-value" is a distribution of probabilities that an observed result is likely to occur by chance, assuming that there is no difference between data subsets.

100. In their field, statisticians accept a p-value below .05 as determining statistical significance. In other words, they accept a 5% error rate. They do not accept a p-value above .05 as statistically significant. This reflects a consensus professional judgment that an error rate above 5% is too great for statisticians to accept an observation as significant proof of an hypothesis.

101. P-values greater than .05 but well below 1.0 may still indicate that an observed difference did not occur by chance. A p-value of, for instance, .20, means that the error rate is 20%. It does not mean that the data lack all meaning, or that a court should exclude the data from consideration as part of a larger body of evidence.

102. The adverse impact ratio ("AIR") is a calculated observed statistic, such as a difference between average performance rates or the difference between promotion rates. Unlike the p-value, it is not a statistical test but is computed from observed samples.

103. The Fisher's Exact Test ("FET") compares the difference between two promotion rates to obtain an observed outcome, then computes a distribution of probabilities that the observed outcome would occur.

104. Both Dr. Silva and Dr. Wiesen analyzed the average performance differences between White and minority police officers at the department level. They agree that excluding Boston, that in 2005 there were statistically significant average performance differences between minorities and White candidates. Dr. Silva and Dr. Wiesen differed in their results for 2007. Dr. Silva did not find a statistically significant pattern of performance differences at the department level, but Dr. Wiesen did. The principal difference is that Dr. Wiesen included departments that did not make promotions. The court accepts Joint Exhibits 133-144 and 153-156 as accurate calculations of the relevant results, using Dr. Wiesen's methodology which the court finds persuasive, with the caveats stated above. The meta-analysis that Dr. Silva and Dr. Wiesen did measured whether the performance differences favored minorities or White candidates. It did not consider promotion rate data. When the candidate pool is small and the selection ratio is small, then it may not be possible to calculate, with statistical certainty, whether average performance differences no longer relate to or impact the AIR for promotions.

B. Methods of Addressing Small Sample Size

105. Small sample sizes can make it hard to draw conclusions from the data. When looking at small populations, adverse impact ratios, standing alone, can be unstable and misleading. For example, if a department has only one minority candidate and that minority is promoted the adverse impact ratio could be 0.0, whereas if that minority was not promoted, the adverse impact ratio could be 1.0. Researchers have found that adverse impact ratios can be an unstable test with samples as large as 200-400. Other factors can influence the appearance of adverse impact, as measured by adverse impact ratios. Those factors include low selection rates (the percent of applicants who are promoted) and a low percentage of minority representation within a jurisdiction. In Massachusetts, selection rates for police sergeant promotions are typically low.

106. There are accepted statistical approaches to analyzing small sample sizes. One of those approaches, employed by Dr. Silva, is to disregard the small samples altogether in the absence of statistical significance. That approach has the virtue of considering only results that meet the professional definition of statistical significance. It has the vice of disregarding large amounts of data that have probative value, particularly when viewed in the context of other corroborative data.

107. Dr. Wiesen employed other methods, including aggregation among departments and across years. He also considered not only statistical significance, but also whether calculations, though not statistically significant by themselves, comprised a body of evidence pointing to a conclusion that could warrant a fact-finder in finding that a conclusion is more likely true than not true. These methods generate higher confidence levels, although they can introduce bias and measure the relevant effect less directly than analysis of unaggregated data.

108. Aggregation of departmental promotional data can introduce bias. For instance, Departments with no minorities cannot have any adverse impact in promotions. Aggregation of candidates from nondiverse departments with diverse departments will impact the White promotional rate from the diverse departments only.[7] Where the aggregated White promotional rate is inflated due to higher White promotion rates from nondiverse departments, that causes the adverse impact ratio to be inflated. This concern does not affect calculations about the impact of the test itself such as difference in scores and passing rates.

109. Combining candidate level data across departments and across years to find a pattern can create bias. A failure to account for repeat test takers may violate the assumption of independence of observations, can also create bias. So can failure to control for the variable selection ratios within each year.

110. The court rejects Dr. Silva's position that it should disregard entirely any results that are not based upon a single test in a single department, after excluding all departments that had no diversity or made no promotions. Dr. Silva's approach is biased in favor of finding no difference in treatment of White, Black and Hispanic candidates. The court agrees that it should give lower weight to calculations that do not meet the criteria Dr. Silva has set forth. However, it does not disregard such results entirely. For one thing, the court must consider the evidence as a whole, giving each part of the evidence the weight it deserves. For another, a strict requirement of statistical significance discards evidence entirely based upon an item-by-item error rate exceeding 5%. But such evidence may have value when combined with other

[7] Aggregating departments that made no promotions with departments that made promotions alters the promotion ratio because it adds candidates who had no chance to be promoted to the denominator of all candidates not promoted. This introduces Statistical Bias in both the white promotional rate and the minority promotional rate, which are compared to calculate an adverse impact ratio.

evidence Under the preponderance of evidence test, the plaintiffs need to prove their case, more likely than not, upon the evidence as whole. Moreover, Dr. Silva's restrictive approach does not follow the Uniform Guidelines, to which the court now turns for guidance in evaluating the probative value of the statistical evidence. The court, of course, recognizes that the Uniform Guidelines are not binding.

C. Guidelines for Assessing Statistics on Adverse Impact

111. The Uniform Guidelines do not call for a statistically significant showing when investigating the existence of adverse impact. Instead, they establish a rule of thumb known as the "four-fifths rule." Under this rule, "[a] selection rate for any race...which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded ..as evidence of disparate impact." 29 C.F.R. § 1607.4(D). Thus, an AIR of 0.80 or less is regarded as evidence of adverse impact.

112. Under the Guidelines (Section 4.D), an adverse impact ratio that is above 0.80 but below 1.0 may still indicate adverse impact if the data are statistically significant:

> A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact. Smaller differences in selection rate may nevertheless constitute adverse impact, where they are significant in both statistical and practical terms or where a user's actions have discouraged applicants disproportionately on grounds of race, sex, or ethnic group.

While the Uniform Guidelines are not binding, the court accepts them as a true statement of what constitutes "evidence of adverse impact for purposes of requiring proof that an exam is a valid measure of job performance."

113. AIR may trigger the four-fifths rule for p-values greater than .05 when samples are small. Small sample sizes may result in unstable results. In such cases, a Shift of One

analysis is an accepted stability analysis to determine whether the observed AIR of a small sample size should trigger validity analysis.

3. Adverse Impacts

a. Total Score and Other Impacts

114. The 2005 statewide examination shows a statistically significant adverse impact on minority test takers with respect to the multiple-choice test score, the Civil Service grade, and the delay to promotion. The mean score difference on the multiple-choice examination was 5.08. The mean score difference on the overall Civil Service grade was 2.01. The mean difference in delay to promotion for minority test takers was -0.89, indicating that it took minority test takers longer to be promoted on average.

115. The Boston 2005 examination shows a statistically significant adverse impact on minority test takers with respect to the multiple-choice test score and the Civil Service grade. The mean score difference on the multiple-choice examination was 6.76, a difference which is "highly practically significant." Meanwhile, the mean score difference was 2.75 on the overall Civil Service grade. Both of these measures had a probability value of well below .001, indicating that the differences are highly statistically significant.

116. The adverse impact ratio for the passing point was 0.60, with 71.7% of White applicants having passed the exam compared to 43.3% of minority applicants, and the adverse impact ratio for promotions was 0.27, with 15.0% of White applicants having been promoted compared to 4.0% of minority applicants. The adverse impact ratios for the passing point and for promotions both fail the federal 80% rule of thumb, and both of these measures were highly statistically significant, with probability values well below .001.

117. The 2006 statewide sergeant examination showed a statistically significant adverse impact on minority test takers on the multiple-choice portion of the Civil Service examination. The mean score difference on the multiple-choice examination was 3.98, a difference causing a lower passing grade and promotion rate among minority candidates. The multiple-choice mean score difference had a probability value of well below .001, indicating that the difference is highly statistically significant. On the 2006 statewide exam, the mean score differences on the overall Civil Service grade also evidence disparate impact on minority test takers, but the differences were not statistically significant.

118. As with the 2006 examination, the 2007 statewide sergeant examination showed a statistically significant adverse impact on minority test takers with respect to the multiple-choice portion of the Civil Service examination. The minority-White mean score difference on the multiple-choice test was 4.46, a difference which is practically significant. The multiple-choice mean score difference had a probability value of well below .001, indicating that the difference is highly statistically significant. The mean score difference on the overall Civil Service grade was 2.75, showing adverse impact on minorities, and was also statistically significant.

119. The adverse impact ratio for the passing point was 0.82, with 68.2% of White applicants having passed the exam compared to 56.2% of minority applicants, and the adverse impact ratio for promotions was 0.36, with 11.5% of White applicants having been promoted compared to 4.1% of minority applicants. The adverse impact ratios for the passing point and for promotions were both less than parity, and the promotion adverse impact ratio fails the 80% rule of thumb. Both of these measures were statistically significant.

120. The 2008 statewide examination also evidenced a statistically significant adverse impact on minority test takers with respect to the multiple-choice portion of the Civil Service

examination. The mean score difference on the multiple-choice examination was 3.32. The mean score difference on the overall Civil Service grade was 0.35. Notably, there was no mean score difference for time to promotion because HRD did not report any promotions of minorities from the 2008 statewide exam.

121. On the 2008 Boston exam, there was statistically significant adverse impact on the multiple-choice examination score and the overall Civil Service grade. The mean score differences were 6.90 for the multiple-choice test and 4.24 for the overall Civil Service grade. Each of these measures had a probability value well below .001, indicating that the differences are highly statistically significant. There was no adverse impact data available with respect to the time to promotion variable because HRD did not provide complete data for promotions from 2008 examination.

122. The adverse impact ratio for the passing point on the 2008 Boston exam was 0.81, with 93.5% of White applicants having passed the exam compared to 75.6% of minority applicants, and the adverse impact ratio for promotions was 0.05, with 9.1% of White applicants having been promoted compared to 0.5% of minority applicants. The adverse impact ratios for the passing point and for promotions were both highly statistically significant, with probability values well below .001.

123. On the 2010 statewide exam, there was statistically significant adverse impact with respect to the multiple-choice test score, which had a mean score difference of 2.63 points.

124. On the 2012 statewide exam, there was statistically significant adverse impact on minority test takers with respect to the multiple-choice examination component. The multiple-choice test had a mean score difference of 3.96 points.

b. Passing Rate Impacts

125. On the 2005 statewide examination, the adverse impact ratio for the passing point was 0.63, with 50.6% of White applicants having passed the exam compared to 31.8% of minority applicants. These figures fail the 80% rule of thumb in the Uniform Guidelines and are statistically significant.

126. On the 2006 statewide exam, the adverse impact ratio for the passing point was 0.79, with 73.7% of White applicants having passed the exam compared to 58.4% of minority applicants. The adverse impact ratios for the passing point fails the federal 80% rule of thumb, and was statistically significant.

127. For the 2008 statewide exam, the adverse impact ratio for the passing point was 0.84, with 81.5% of White applicants having passed the exam compared to 68.9% of minority applicants, and was statistically significant. While the 2008 statewide passing point adverse impact ratio passes the federal 80% rule of thumb, it still indicates adverse impact because the data are statistically significant.

128. On the 2010 statewide exam, the adverse impact ratio for the passing point was 0.51, with 17.6% of White test takers having passed and 8.9% of minority test takers having passed. That adverse impact ratio of 0.51, indicated that minorities failed the exam at twice the rate of non-minorities. That was less than the federal 80% rule of thumb. Significantly, only half the proportion of minority test takers were potentially eligible for promotion compared to White test takers.

129. On the 2012 statewide exam, the adverse impact ratio for the passing point was 0.64, with 36.7% of White test takers having passed and 23.5% of minority test takers having passed. That ratio is less than the federal 80% rule of thumb, and was statistically significant.

Again, this adverse impact ratio is important because only half the proportion of minority test takers were potentially eligible for promotion compared to White test takers.

c Effects on Promotion: Rate of Promotion and Delay in Promotion

130. On the 2005 statewide exam, the adverse impact ratio for promotions was 0.22, with 14.0% of White applicants having been promoted compared to 3.0% of minority applicants. The adverse impact ratios for the passing point and for promotions both fail the federal 80% rule of thumb. Both of these measures were statistically significant.

131. In 2005, there were statistically significant average performance differences in promotions between minority officers compared with White officers using a two-tailed p-value. Both Dr. Silva and Dr. Wiesen agreed on that point.

132. On the 2006 statewide exam, the adverse impact ratio for promotions was 0.18, with 14.3% of White applicants having been promoted compared to 2.6% of minority applicants. The adverse impact ratio for promotions fails the federal 80% rule of thumb, and was statistically significant.

133. On the 2006 statewide exam, the delay to promotion evidenced disparate impact on minority test takers, but the differences were not statistically significant.

134. In 2007, including departments that made no promotions, there were statistically significant average performance differences between minority officers compared with White officers on the statewide sergeant promotional exam.

135. From the 2008 statewide exam, the adverse impact ratio for promotions for the 2008 statewide examination was 0.0, with 2.9% of White applicants having been promoted and 0.0% of minority applicants having been promoted.

136. In 2009, HRD delegated promotions to municipalities and as a result, HRD did not have promotional data for the 2010 and 2012 statewide promotional exams.

d. Impacts in Municipalities Outside Boston

137. Impacts in individual municipalities also occurred outside Boston. For instance, consider ng p-values greater than .05 proves, more likely than not, that an adverse impact also occurred in Springfield. The (two-tailed) p-value for the rate of promotion of minority candidat;s relative to White candidates within the Springfield Police Department for the 2005 examinaiion was .07. There is "a 93 percent chance that [the difference in promotion rates between minority and White candidates] is due to a lack of equality in ... promotion[.]" Similarly, the p-value for the promotion rate for the 2007 examination in Springfield was 0.26, indicatin g that there was a 74%chance that the difference in promotion rates was due to lack of equality n promotion. Notably, however, aggregating the 2005 and 2007 Springfield examinations, neither of which individually resulted in enough promotions to yield a statistically significant result, results in a statistically significant p-value of .04, suggesting that there was a 96%chance that the difference in minority and White promotion rates was due to something other than chance.

138. The same analysis for the MBTA Transit Police showed a pattern that was very consistent with the overall lower promotion rate for minorities than for nonminorities."

139. Statistically significant mean score differences individually within several municipalities existed between 2005 and 2012. Specifically, Dr. Wiesen found statistically significant mean score differences favoring Whites over minorities in New Bedford, Randolph, and Sprir gfield in 2005, in Brockton, Lawrence, and Lowell in 2006, in Brockton and Holbrook in 2008, in Cambridge in 2010, and in Lawrence and Newton in 2012.

140. This pattern, extending at least forty years into the past, of low minority passing and promotion rates, is sufficient evidence of adverse impact to require HRD to produce evidence of the validity of its examinations.

Adverse Impact Across Years

141. The statewide and Boston sergeant examinations also demonstrate adverse impact against minority test takers when results are aggregated across the years between 2005 and 2012.

142. Aggregation of data over time has the advantage of showing "the big picture" regarding the disparate impact of HRD's format.

143. Combining candidate level data across departments and across years to find a pattern also has a disadvantage, because repeat test takers can violate the assumption of independence of observations, and because selection ratios vary within each year. Moreover, candidates do not compete with every other candidate across the state for promotions. They do not need to be in the top 15%to be promoted, in fact, promotions were made at the 30th, 36th and 54th percentiles of candidates, ranked by score percentile.

144. HRD did not provide the information required to identify repeat test takers. It was therefore impossible to remove repeat test takers from the data. However, repeat test takers may well generate sufficiently independent observations if they changed their preparation strategy, or prepared more, or if the passage of time gave them significantly more experience upon which to draw.

145. Rather than ignore aggregation of data over time, the court considers this evidence, albeit with caution and with the knowledge that there may be some statistical bias in the results.

146. To aggregate data over time, Dr. Wiesen evaluated the rank order placement of White and minority candidates statewide between 2005 and 2012, and in Boston for 2005 and 2008. He calculated a percentile score for each applicant within each exam, ranging from 0 to 100 for each exam. He then grouped those percentile scores for all examinations and evaluated their distributions by ethnic group. The percentile scores were grouped in increments of 5% (i.e., the highest scoring 5% of test takers, the next highest 5% scoring test takers, etc.).

147. Police departments make selections for promotions from the top candidates in rank order based on what is known as the 2N+1 rule.[8] Thus, if relatively few minority test takers appear in the top 5% or 10% of test takers, it is far less likely that a minority candidate will be promoted off of the eligibility list, and far more likely that minority candidates will experience a delay in promotion compared to White candidates.

148. Both statewide and for Boston, ranking the scores by percentiles over time demonstrates a statistically significant adverse impact on minority test takers. That is, a clear pattern of relatively fewer minorities falling in the top 5% or 10% of all test takers compared to White test takers.

149. The scores between 2005 and 2012 show a pattern: fewer minority test takers scored in the top 5% of scores or in the next highest 5% of scores. Fewer than 1% of minority test takers (or 3 individuals) scored in the top 5% of all test takers compared to 5.2% of White test takers (or 194 individuals), and only 2.8% of minority test takers (or 11 individuals) scored in the next highest 5% of scores compared to 5.6% of White test takers (or 209 individuals).

[8] For example, if there are three vacancies in a given municipality, the three candidates selected for promotion are chosen from among the top seven candidates on that municipality's eligibility list. See G.L. c. 31 § 27.

150. If scores were equally distributed, there would have been approximately 19 minority test takers (compared to 3) in the top 5% of scores and 22 minority test takers (compared to 11 in the next 5% of scores.

151. This disparity between minority and White applicants is reversed for the lowest scoring 5% of test takers. Approximately 11.0% of minority test takers fell into the lowest scoring 5% of test takers, but only 4.3% of White test takers fell into this grouping.

152. The Boston results display a similar pattern of statistically significant adverse impact when test takers are aggregated across years. Fewer minority test takers scored in the top 5% of scores or in the next highest 5% of scores for the two Boston examinations combined. Fewer than 1% of minority test takers (or 4 individuals) scored in the top 5% of all test takers compared to 7.6% of White test takers (or 50 individuals), and only 1.8% of minority test takers (or 8 individuals) scored in the next highest 5% of scores compared to 7.0% of White test takers (or 46 individuals).

153. An equal distribution in the top two groupings for Boston would have approximately 22 minority test takers (compared to 4) in the top 5% of scores and 22 minority test takers (compared to 8) in the next 5% of scores.

154. This disparity between minority and white applicants in Boston is reversed for the lowest scoring 5% of test takers. Approximately 8.7% of minority test takers fell into the lowest scoring 5% of test takers, but only 2.4% of White test takers fell into this grouping. In general, the lowest eight percentile groupings had relatively more minority test takers and fewer White test takers, while the eight highest percentile groupings had relatively more White test takers and fewer minority test takers.

Historical Pattern of Disparate Impact

155. The Uniform Guidelines (Section 4.D) also contemplate establishing adverse impact by looking at historical patterns:

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

156. Even for a jurisdiction where numbers are too small to obtain a statistically significant measure of adverse impact, an historical pattern, over three or more years, reflecting low rates of minority promotions, would suffice to require evidence of test validity.

157. Joint Exhibit 266 documents a long pattern of HRD administering examinations that have had a disparate impact on minority candidates within Boston and the Commonwealth.

158. In 1970, the City of Boston's written police officer examination resulted in passing rates of 25% for Black candidates and 10% for Spanish-surnamed candidates, compared to 65% of White candidates. See Castro v. Beecher, 334 F. Supp. 930, 942 (D. Mass. 1971). This resulted in passing rate adverse impact ratios of 0.38 for black candidates and 0.15 for Spanish-surnamed candidates.

159. In 1974, HRD administered an examination for the Boston Police Department in which only 8% of Black police officers who took the exam (comprising only 2 individuals) were promoted to sergeant, while 17% of White officers who took the examination (comprising 104 individuals) were promoted to sergeant. That examination had a passing point adverse impact ratio of 0.62.

160. In 1977, HRD administered another police sergeant examination for the Boston Police Department. That examination had high adverse impact at the passing rate, with only

4.5% of Black test takers passing compared to 16% of White test takers, resulting in a passing point adverse impact ratio of 0.28.

151. In 1985, Boston and HRD developed and administered a validated examination consisting of multiple components with a low passing point adverse impact ratio of 0.85, which passed the 80% rule of thumb. However, two years later, Boston reverted to its old format, resulting a passing point adverse impact ratio of 0.5 on the 1987 examination.

152. According to HRD's 1991 Validation Report, Boston's 1991 examination had adverse impact on minority test takers at two contemplated passing points: 0.16 at 70 points; 0.34 at 60 points.

153. The 1991 Validation Report also analyzed mean scores by racial/ethnic group. The mean score of minority candidates who took the 1991 examination was considerably lower than that of White candidates: the mean score difference between White and Black candidates was 11.8 points, and the mean score difference between White and Hispanic candidates was 9.1 points.

154. The 1996 sergeant examination had adverse impact on Black and Hispanic candidates. Specifically, passing rates for Black and Hispanic candidates were, respectively, 22.73% and 26.67%, compared to 53.71% for White candidates. These corresponded to passing point adverse impact ratios of 0.42 for Black candidates and 0.5 for White candidates.

155. This pattern of historical adverse impact is clear at the named plaintiffs' individual municipality-level as well. There were no minority police sergeant promotions in Worcester for 14 years. There were no minority police sergeants in Brockton from approximately 2000 to 2012.

III. Alternatives with less adverse impact

166. Many alternatives would have less impact than a rank-order list based upon a rote-memory multiple choice test with 80 questions.

167. Because unnecessarily large cognitive loads tend to create the most adverse impact, one class of alternatives involves reducing the cognitive load.

168. Even without changing the current test format of a multiple-choice examination plus an E&E component, HRD could reduce adverse impact by having skilled test developers design questions that avoid rote memory answers and instead test situational judgment and other types of skills. Writing questions in plain language, instead of using convoluted phrases, likely would have a similar beneficial effect.

169. A simple way to reduce the cognitive load is to use fewer questions. Using an exam with 35 questions, rather than 80 questions, reduces that cognitive load and has been shown to have less adverse impact.

170. Shortening the reading list would also reduce the cognitive load. It was not necessary, for instance, to include a 600-page Police Administration textbook on the 2008 reading list, when the exam included only 2 questions from that book on the statewide exam and 1 question on the Boston exam.

171. While knowing certain information is necessary to be a good sergeant, HRD could ensure an adequate base of knowledge by scoring the written exam on a pass-fail basis. HRD or appointing authorities could then assess the untested KSAs through interviews, comprehensive review of past accomplishments or other methods to test and score a candidate's key leadership abilities. If doing so were deemed too expensive at the statewide level, it could provide a rank list to appointing authorities, leaving it to local chiefs to assess those qualities.

172. All of the above alternatives are inexpensive. Using fewer multiple choice questions may even save some test development costs.

173. Another alternative is banding, which "recognizes error in measurement and creates a range of observed scores that are functionally the same." Letter of March 30, 2009, from Dr. Rick Jacobs, CEO of E.B. Jacobs, to Personnel Administrator Dietl ("March 2009 Letter") regarding "promotion test scoring and the use of score banding to determine candidate qualifications.").

174. It is not necessary for HRD to rank candidates based upon a single point difference between candidates. Banding scores within a range would make more candidates eligible for promotion. The "tie-breaker" could be interviews, comprehensive review of past accomplishments or other methods to test the key leadership abilities.

175. In late 2008 and 2009, HRD decided to create 11-point bands both statewide and in Boston based on a report from E.B. Jacobs, for whom Dr. Silva was working at the time.

176. Dr. Jacobs' March 2009 letter pointed out that score banding systems are common and that '[i]n schools we have bands but call them grades" such that, for instance, a score of 93, 95 or 97 would all receive an "A." He stated:

> I recommend the use of banding, because by banding promotional test scores we are (1) recognizing that there is measurement error, (2) using the level of error to determine the actual width of the band and determining candidates equivalent within that level of error, and (3) creating a pool of candidates, those with a band, who will be seen as equally qualified based on their test scores so that the use of another variable or variables will be necessary to make the final decision among those individuals. In the context of police and fire promotional testing candidates have long and important job performance records upon which they can be judged. By relying on only a test score much of the contributions a candidate has made to a department and many of the abilities/competencies they have developed relating to the next level job may not be considered in the promotional process. With banding candidates who are equally qualified based on exam performance can then be further considered based on other important job relevant characteristics.

177. HRD received a banding proposal from E.B. Jacobs on January 21, 2009. At least internally, HRD's Director of its Organizational Development Group reported that "According to our experts, banding is considered a best practice. The scheme that they have developed for us is based on scientific testing standards and is valid and defensible."

178. Dr. Jacobs' March 2009 Letter and HRD's own proposals and internal writings demonstrate that it knew by late 2008 or by 2009 that its police promotional examinations lacked sufficient validity to be used in rank order fashion, without banding, despite using them in such fashion for all in-basket exercises and incident command exercises of the years at issue in this case.

179. The addition of components designed to test skills other than technical knowledge results in much greater validity in an examination process. Methods that increase validity and decrease adverse impact include assessment centers, structured oral interviews, written exercises, career boards, in-basket exercises and incident command exercises, and tactical exercises.

180. Subordinate role-play or a citizen meeting or group meeting may improve adverse impact, depending upon the weight of the components, the other components, and the different knowledges and areas that the overall exam covers.

181. Other components of a police promotional exam, such as oral assessment centers exercises that focus on how a candidate might orally respond to a situation, can have significantly less adverse impact.

182. The assessment of certain nonwritten skills, such as leadership, conscientiousness, calmness under pressure, decision-making, interpersonal skills, and oral communication tend to have low or no adverse impact. As Ms. Fitzgibbons testified, these skills cannot be tested on a

multiple-choice rote memory examination, but can be tested through additional assessment components.

183. There is ample scholarly literature in the field of industrial psychology stating that adverse impact can be significantly lowered by utilizing additional assessment exercises that more closely mirror the actual duties of the job. The literature also shows that assessment centers increase validity because they increase the representativeness and the fidelity of the examination processes.

184. These techniques are in use including by experts in this case including Dr. Ralfilsor, Dr. Silva and his firm E.B. Jacobs, and Dr. Outtz.

185. Prior to the exams in this case, HRD and municipalities under its jurisdiction proposed alternatives to reduce adverse impact.

186. The 1985 Boston exam included an assessment center component in addition to multiple-choice and E&E components. The passing point adverse impact ratio for minority candidates on the 1985 Boston exam was .85.

187. The 2002 Boston sergeant's exam, which included a structured interview component in addition to multiple-choice and E&E components, had a promotion rate adverse impact ratio for Black candidates of .327 with the structured interview component, and an adverse impact ratio of .322 without the structured interview component. The 2002 Boston exam had a promotion rate adverse impact ratio for Hispanic candidates of .270 with the structured interview component, and an adverse impact ratio of .133 without the structured interview component.

188. The 2002 Boston exam with the structured interview component resulted in one additional promotion of a Hispanic candidate, and no additional promotions of Black candidates, compared to the exam without structured interview component.

189. HRD has argued that development of small, medium, and large department job analyses would require HRD to administer four different job analyses could take 6 months to a year. It contends that a statewide assessment center would be a massive logistical effort in a very short period of time that would require multiple consulting firms to accomplish. These cost considerations apply to some, but not all, of the alternatives discussed above. Whatever the law on consideration of cost, and whatever HRD's budgetary constraints, there are alternatives, discussed above, that have no or minimal cost impact.

190. Moreover, HRD has overstated the costs involved. One study, discussed in Lopez I, found that the utility of an assessment center ranged from about $500 to $3,000 per candidate. More significantly, HRD's cost concerns fail to account for cost-saving strategies. For instance, it is common to implement work simulations, roleplaying exercises, and subordinate exercises via video presentation rather than live in-person.

191. Even if some multi-component processes might impose some additional cost, they also result in benefits, including the appointment of more qualified and diverse police sergeants. Having more qualified supervisors can avoid other financial costs, such as lawsuits alleging civil rights violations. Intrinsic benefits also resulting from greater minority representation in the ranks of police sergeants by, for instance, generating a corps of sergeants that reflects the communities they are policing and creating a greater pool of minorities for higher level positions.

IV. HRD's Knowledge

192. The above also establishes HRD's knowledge, for years before and during administration of the challenged tests, of its format's adverse impacts, the tests' substantial lack of job-relatedness, and the availability of alternative methods with less adverse impact. The court supplements that discussion with some additional findings.

193. Historical materials in evidence, including case law, verified complaints, and expert reports relating to prior to the exams at issue, as well as an HRD 2006 report pertaining to the 2005 Boston exam results prove HRD's awareness of the disparate impact of its practice of creating rank-order lists generated by rote-memory multiple choice exams. Indeed, HRD acknowledged and tried to remedy these disparities prior to the exams in this case. It has known for over 50 years that the format for the written portion of its police sergeant promotional examinations was not based upon any valid scientific analysis or assessment of the skills and abilities necessary to perform well as a police sergeant.

194. Beginning in the 1970s, a series of lawsuits, court decisions, and expert and self-analysis by HRD proves that HRD knew that its written multiple-choice police sergeant promotional examinations had an adverse impact on minority candidates and were not sufficiently job-related as to be valid measures of job qualifications. Indeed, HRD's own internal reports and records dating back to at least 1991 demonstrate that HRD has long possessed such knowledge.

195. The data established that minority police officers who took the 1974 and 1977 Boston Police Department sergeant's examinations suffered statistically significant adverse impact. For example, 8% of Black police officers who took the 1974 Boston Police Department sergeant's examination (comprising only 2 individuals) were promoted to sergeant, while 17% of

White officers who took the examination (comprising 104 individuals) were promoted to sergeant. The 1977 Boston Police Department sergeant examination had similar adverse impact at the passing rate, with only 4.5% of Black test takers passing compared to 16% of White test takers.

196. In 1985, in connection with a 1980 consent decree, HRD and the City of Boston utilized an outside expert consulting firm to develop a valid multi-component exam for Boston. Plaintiffs' expert, Dr. Wiesen, who was then working for the Commonwealth as an industrial psychologist, oversaw this process in part. According to a letter by the then-personnel administrator, David Haley, the process was successful, as the examination had a passing point adverse impact of only .85, which resulted in many Black candidates becoming eligible for promotion.

197. In 1988, the Civil Service Commission ruled that the written multiple-choice job knowledge test, even when combined with the E&E component, was not a fair or reliable test of the skills and abilities necessary for the job of police lieutenant because it could not assess one of the most important aspects of the job: supervisory ability.

198. As reported in HRD's 1991 Validation report, the adverse impact ratio for the 1991 Boston Police examination for the passing score of 70 was 0.16. The 1991 Validation Report also analyzed mean score differences between minorities and Whites. The mean score difference between White and Black candidates was 11.8 points, and the mean score difference between White and Hispanic candidates was 9.1 points. These results were statistically significant.

199. In 1992, HRD administered another sergeant's promotional examination both for Boston and statewide which had significant adverse impact. As a result of the 1992 examination,

the City of Boston, with HRD's approval, sought to promote a number of minority candidates to the position of sergeant out of order, contending that they were equally qualified and that the failure to promote candidates out of order would result in significant adverse impact. The Civil Service Commission, however, held that departure from strict rank order was unauthorized.

200. A number of experts have warned HRD that its police sergeant promotional examinations were not valid, particularly when used to make selections in rank order fashion. HRD agreed that without banding, the test results would continue to have a disparate impact on minority candidates.

201. In 1996, HRD administered another multiple-choice sergeant's examination statewide and within Boston. Both the examinations produced statistically significant adverse impact. Cotter v. City of Boston, 193 F. Supp. 2d 323 (D. Mass. 2002), aff'd in part and rev'd in part Cotter v. City of Boston, 323 F.3d 160 (1st Cir. 2003)

202. Accordingly, the MBTA, with HRD's approval, used a separate minority promotion list in order to promote several minority candidates. As of 1996, minorities represented 27.6% of patrol officers in the MBTA police (fifty officers out of 181), but only one sergeant out of fourteen was a minority. If the 1996 promotions had been made in strict rank order, no Black candidates would have been promoted and the proportion of minority sergeants would have dropped to only 4% (i.e., one minority sergeant).

203. In 2006, HRD issued a report discussing and analyzing results of the 2005 police sergeant promotional examination for the Boston Police Department. According to the report, the mean score difference for White and Black candidates was 7.09, which was statistically significant. The passing point adverse impact ratio between Black and Hispanic candidates on the one hand, and White candidates on the other, was 0.59, which fails the federal 80% rule of

thumb. The report concluded that, "Overall race appears to be a factor in performance results with White applicants outperforming Black, Hispanic, and Native American applicants."

204. In 2006, HRD's statewide police sergeant promotional examination had a passing point adverse impact ratio of 0.75, again failing the federal 80% rule of thumb. HRD reported the pass-fail results of that examination, indicating its awareness of adverse impact.

205. In 2008, HRD summarized promotional data from the 2006 statewide examination. All of the individuals promoted from that examination were White, resulting in a promotion adverse impact ratio of 0.0.

206. Cities throughout the Commonwealth submit so-called "Form 67s", which list the number of minorities in various public safety positions, to HRD every year. The Form 67 for the city of Brockton demonstrated that despite its large minority population there were no minority police sergeants in Brockton in 2006.

207. Between approximately 1987 and 2001—a period of 14 years—there were no minority police sergeant promotions in Worcester. In the Brockton Police Department from 1996 to 2019 there were no minority police sergeants from approximately 2000 to 2012.

208. Thus, when it administered and scored the exams at issue, HRD unequivocally knew that rote-memory multiple choice exams that generate a list, ranked in order of scores, systematically affects Black and Hispanic candidates adversely compared to White candidates. It knew that its rank order lists and that administration and scoring had a significant disparate impact on Black and Hispanic police officers seeking promotion to sergeant. It knew that this impact reduced the number of promotions to sergeant in large departments and was likely to do so in smaller departments even though it is very difficult to realize in which small department(s) the impact would occur in any given year.

DISCUSSION

Section 4(4A) ("Section 4A") of G.L. c. 151B prohibits "interfere[nce] with . . . the exercise or enjoyment of any right granted or protected by this chapter," including the right to be free from discrimination in the terms, conditions, and privileges of employment. At issue, in this case, are claims of interference with the right to equal opportunities for promotion without discrimination on the basis of race, color, or national origin.

The word "interfere" "implies some form of intentional conduct," but does not require "a specific intent to discriminate." Lopez II, 463 Mass. at 708-709.

> [A]n interference claim under [G.L. c. 151B,] § 4 (4A) may be established by evidence of disparate impact. Because discrimination based on proof of disparate impact does not require proof of discriminatory intent, the element of intentionality is satisfied where it is shown that a defendant knowingly interfered with the plaintiffs' right to be free from discrimination in the terms, conditions, and privileges of employment on the basis of a protected category such as race, color, or national origin. Thus, to make out a prima facie claim under § 4 (4A) based on a disparate impact theory of liability, a plaintiff must allege facts that, if proved, would establish that (1) a defendant utilized specific employment practices or selection criteria knowing that the practices or criteria were not reasonably related to job performance; and (2) a defendant knew that the practices or criteria had a significant disparate impact on a protected class or group.

Id. 463 Mass. at 711. Based upon the court's findings set forth above, the plaintiffs have proven that, more likely than not, HRD has interfered with their rights to an equal opportunity for promotion to police sergeant without racial or national origin discrimination.

I. Knowing Use of Practices Not Reasonably Related to Job Performance

a. Disparate Impact

First, the plaintiffs must show that the challenged employment practice had a significant disparate impact "on promotional opportunities for employees of a particular race, color, or national origin." Lopez II, 463 Mass. at 709. Disparate impact "involve[s] employment

practices that are facially neutral in their treatment of different groups, but that in fact fall more harshly on one group than another." Id. (citation and internal quotations omitted).

HRD's format had a significant disparate impact on Black and Hispanic police sergeant promotion candidates on three separate grounds:

> (1) There is statistically significant proof of disparate racial and national origin impact on promotions in a number of municipalities where promotions were made and there were Black or Hispanic candidates or both.
> (2) The court infers impact on promotions from the statewide disparity in scores and passing rates, coupled with the lack of job-relatedness.
> (3) There are data sets that lack statistical significance considered separately, but point consistently to adverse impact on promotions of Black and Hispanic candidates

While each of these grounds independently supports this court's inference of adverse impact, they powerfully reinforce each other when considered together. See Smith v. City of Bos., 144 F.Supp.3d 177, 194 (D. Mass. 2015) ("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."); Bradley v. City of Lynn, 443 F.Supp.2d 145, 158 (D. Mass. 2006) (considering multiple data points beyond mere hiring rates).

The court now turns to each of these grounds in more depth.

First, some results are statistically significant, standing alone. In 2005 and 2008, the City of Boston had statistically significant adverse impact in promotions of Black and Hispanic officers. On the 2005 statewide test, there were statistically significant average performance differences between minority officers compared with White officers using a two-tailed p-value. Both Dr. Wiesen and Dr. Silva found statistically significant disparate impact in minority-White performance on several of the examinations at issue in this case at the municipal level including Boston, Randolph, Springfield, and Brockton. It is likely no accident that these statistically significant results occur in large sample sizes. Where large data sets permit high confidence

calculations, the inherent tendency of HRD's exams to cause adverse impact becomes statistically certain. The absence of statistical significance in other data sets most likely reflects the sample size, rather than the absence of embedded racial and national origin discrimination. The other evidence in this case persuades the court that HRD's format interfered with the plaintiffs' rights in contexts that did not meet strict criteria for statistical significance.

Second, disparate impacts at certain stages of the selection process support an inference of adverse impact on promotions of Black and Hispanic candidates.

The Supreme Judicial Court has recognized the inference that may flow from disparities in examination results, coupled with proof that the exam does not predict job performance:

> It was not necessary that the plaintiffs allege that use of the division's examination led to a disparate impact on promotions in any particular, identified, employing municipality in order to state an interference claim under § 4 (4A). An allegation that a Statewide examination has been shown to disproportionately disadvantage African-American and Hispanic candidates, and is not a predictor of job performance, implies that use of the examination will have a disparate impact on the employment opportunities of at least some African-American and Hispanic police officers within the Commonwealth, by limiting the number of qualified African-American and Hispanic candidates among whom individual municipalities using the examination might seek to make promotions.

Lopez II, 463 Mass. at 712. As a matter of law, then, the inference of "disparate impact on the employment opportunities of at least some African-American and Hispanic police officers within the Commonwealth" follows from logic and common sense.

The Commonwealth reads this passage as applying only to the pleadings stage, and addressing only the sufficiency of allegations in the complaint. It maintains that, when it comes to trial, the plaintiffs' burden includes proving a disparate impact on promotions in specific municipalities. To be sure, in Lopez II, the Supreme Judicial Court only considered the adequacy of the complaint. But, to withstand dismissal for failure to state a claim, the complaint must set forth "factual 'allegations plausibly suggesting (not merely consistent with)' an

entitlement to relief, in order to 'reflect[] the threshold requirement of [Fed. R. Civ. P.] 8(a)(2) that the 'plain statement' possess enough heft to 'sho[w] that the pleader is entitled to relief.'" Iannacchino v. Ford Motor Co., 451 Mass. 623, 636 (2008), quoting Bell Atl. Corp. v. Twombly, 550 U.S. 544, 127 S. Ct. 1955, 1966 (2007). It would be odd to hold the complaint sufficient, without an alleged disparate impact on promotions in any particular municipality, and then to require proof of that unalleged fact. That reading also seems inconsistent with the Supreme Judicial Court's statement about inferences that may be drawn from statewide analysis, coupled with proof that the exam fails to predict job performance. In any event, the plaintiffs have met their burden of persuasion on either reading of Lopez.

The court draws an inference of discrimination here. The chain of logic is simple: HRD's format produces a rank-order list with an adverse impact in scoring and passing rate. It circulates that list for use by appointing authorities in promoting police officers to sergeant. The employer uses the biased list of test scores, in rank order, to decide who gets promoted. There is no process to purge the list of bias when promotions occur. With an adverse impact in the scoring, passing rate and rank order, an adverse impact upon promotions based upon HRD's list is highly likely. To be sure, it is not always easy or even possible to identify which promotions in which municipalities reflect this bias, but the plaintiffs do not have to prove their case with such specificity.

Third, there are additional facts and calculations that, while not statistically significant in themselves, collectively demonstrate in convincing fashion the adverse impact of HRD's format upon the plaintiff class. This additional evidence largely falls into two, sometimes overlapping, categories: (1) data sets that fail the Uniform Guidelines' four-fifths rule and therefore call for

demonstration of test validity and (2) small data sets that, when considered with other such sets, show a consistent trend of adverse impact.

Based upon adverse impact ratios, there is a long pattern of disparate impact on minority candidates from HRD's examinations. See Brackett v. Civil Service Commission, 447 Mass. 233, 246 (2006) (regarding the discriminatory effects of HRD's examination in the 1990s, "If the MBTA had based its promotion decisions on strict rank order...[n]o black officers would have been promoted to fill any of the seven sergeant positions."). Exhibit 266 details adverse impact ratios below 0.80 for 15 exams from 1970 to 2009 for the passing point and an adverse impact ratio of 0.0 for promotions from the statewide police sergeant exam in 2006. Additional pass-fail adverse impact ratios below 0.80 also occurred in:

- 2010 statewide, accompanied by a statistically significant difference in scores according to one-sided p-value and a two-sided p-value just outside statistical significance (0.066).
- 2012 statewide, accompanied by a statistically significant difference in scores and in pass-fail rates.

By 2008 HRD provided promotional data on a spotty basis and by 2010, no longer made promotional data available, making assessments of adverse impact ratios for promotions impossible. The Court nevertheless finds persuasive evidence under the Uniform Guidelines of both the exams' tendency and the reality of adverse impact. Because HRD knew these results, it also tends to show HRD's knowledge of adverse impact.

Moreover, for the vast majority of scoring, passing and promotion rates that are not statistically significant, there is a consistent pattern: they tend strongly to point in the direction of adverse impact upon Black and Hispanic candidates. For instance, on the 2005 statewide exam,

out of 16 police departments, 14 had scoring differences between groups that favored White candidates over minority candidates. The overall pattern was highly statistically significant ($p = 0.002$). Similar statistically-significant differences occurred on the 2006, 2007, 2008, and 2012 statewide exam, with the 2010 statewide exam falling just outside of statistical significance ($p = 0.072$). Other aggregations within departments and over multiple exam-years tell the same story. For instance, over the years, a disproportionate number of White test-takers fell within the top percentile groupings (measured in approximate 5% increments) compared to minority test takers. The numbers are reversed for the bottom percentiles. This trend holds both for Boston exams and statewide testing. The trend is important, because police departments promote from the top of the list. Disparate impact in promotions is therefore highly likely, even where it is not possible to prove to a statistical certainty that a particular promotion resulted from the unequal distribution.

The Supreme Judicial Court cited aggregate data in Lopez II. See 463 Mass. at 712 n.20, 714 n.23 In Tatum, the Appeals Court stated that the SJC "sanctioned the use of significant Statewide statistics to show disparate impact." 2020 WL 4200865 at *1 n.5. See Bradley, 443 F.Supp.2d at 149 (statistical evidence showed adverse and disparate impact).[9] Whether or not the SJC statements are "dicta," this court is bound to follow what the SJC and Appeals Court say, not just what may narrowly qualify as a holding.

The non-binding Uniform Guidelines also provide assistance in evaluating the statistical evidence They specifically approve of the practice of aggregation:

---

[9] The Commonwealth cites Lopez I for the proposition that "a municipality's promotions should be assessed with respect to the pool of candidates actually available for appointment to rank of sergeant . . . What adverse impact, if any the test might have with respect to another municipality's candidate pool is simply not relevant." 2014 WL 12978866 at *10. When HRD is the defendant, this court disagrees, because aggregate data have probative value in showing adverse impact resulting from a testing bias that operates in all municipalities and likely produces biased promotions in at least some municipalities.

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

29 C.F.R. § 1607.4(D); see also Uniform Guidelines Q&A, 44 Fed. Reg. 11999-12000 (1979) (Q27: "If the test is administered and used in the same fashion for a variety of jobs, the impact of that test can be assessed in the aggregate."). For small data sets, the Uniform Guidelines also support evaluating adverse impact by analyzing patterns of adverse impact over time. See Uniform Guidelines Q&A, 44 Fed. Reg. 11999-12000 (1979) (Q21: "if a lower selection rate continued over a period of time, so as to constitute a pattern, then the lower selection rate would constitute adverse impact...").

To be sure, the relationship between test scores and any given promotion is not a direct one. It depends on many things, including individual candidate performance. It also turns on the department's selection rate, which is the number of vacancies it is seeking to fill through a promotion. Fully aware of these complexities, the SJC held that "employment procedures or testing mechanisms that operate as 'built-in headwinds' for minority groups" can establish adverse impact even absent discriminatory intent. Lopez II, 463 Mass. at 709-710, (quoting Griggs v. Duke Power Co., 401 U.S. 424, 432 (1971)); see also Connecticut v. Teal, 457 U.S. 440, 451 (1982) ("The 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.") (emphasis in original). Even if the disparate impact of HRD's format is one factor among others, it generated precisely that kind of "headwinds" and is an important cause of disparate promotions.

Finally, these three types of statistical evidence have much greater persuasive power when considered together, rather than separately. If there were no statistically significant demonstrations of adverse impact on promotions, there would be greater reason to question an influence of adverse impact from disparate test scores or passing rates. The same is true if there were no adverse impact ratios below 0.80, or if there were no clear trend in the small data sets for individual appointing authorities. But all these types of data point to the same conclusion: HRD's format had a disparate impact upon promotions (and delay in promotions) of Black and Hispanic candidates for the years in questions. The court adopts that conclusion by a preponderance of the credible evidence.

II Lack of Job Relatedness

Second, the plaintiff must demonstrate that the challenged practice is "not reasonably related to job performance." Lopez II, 463 Mass. at 711. This is a matter of degree, as the word, "reasonably" suggests. The parties agree that at least some aspects of the police sergeant's job lend themselves to a multiple-choice test. For instance, written tests are useful to assess practical knowledge used in performing tasks equivalent to those used in a desk job. The plaintiffs have proven, however, that HRD goes well beyond that scope and measures matters that are not reasonably related to job performance.

Following Lopez II, 463 Mass. at 703-704 and 703 n.8, the court looks to federal authority in construing the job-relatedness component of plaintiffs' § 4(4A) claim. Anti-discrimination law "has forbidden giving [selection] devices and mechanisms controlling force unless they are demonstrably a reasonable measure of job performance. . . . What Congress has commanded is that any tests used must measure the person for the job and not the person in the abstract." Albemarle Paper Co. v. Moody, 422 U.S. 405, 426 (1975) (citation and internal

quotation marks omitted); see also Smith v. Boston, 267 F.Supp.3d 325, 333 (D. Mass. 2017) ("a court ensures that a selection device evaluates characteristics important to job performance"); Vanguard Justice Soc'y, Inc. v. Hughes, 592 F.Supp. 245, 258 (D. Md. 1984) (accord) ("In short, an examination is content valid if it tests knowledges, skills and abilities critical to a job and thereby rates applicants on the basis of their ability to perform that job.").

"Evidence of the validity of a test or other selection procedure by a content validity study should consist of data showing that the content of the selection procedure is representative of important aspects of performance on the job for which the candidates are to be evaluated." 29 C.F.R. § 1607.5(B); Accord Smith, 144 F.Supp.3d at 206-207 (accord); Vanguard Justice Soc'y, Inc., 592 F. Supp. at 266. The Court's findings of fact demonstrate the lack of content validity in HRD's format.

Here, the issue is not just the validity of the exam itself, but also of ranking candidates by their numerical scores. Therefore, "evidence 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 under these guidelines ... the user should have sufficient evidence of validity and utility to support the use on a ranking basis." 29 C.F.R. § 1607.5(G). See 29 C.F.R. § 1607.14(B)(6) ("users should evaluate each selection procedure to assure that it is appropriate for operational use, including establishment of cutoff scores or rank ordering"). There must be a relationship between higher scores and better job performance. See 29 C.F.R. § 1607.14(C)(9). Similarly, where a rank order selection procedure includes consideration of prior training or experience as a selection criterion, there must be a correlation between that training and experience and the content of the job. See 29 C.F.R. § 1607.14(C)(6).

Moreover, using a score to rank and then select candidates requires a strong correlation between test scores and job performance. See Brunet v. City of Columbus, 1 F.3d 390, 410 (6th Cir. 1993) (using exam score to rank candidates requires strong correlation between test score and job performance); Ensley Branch of NAACP v. Seibels, 616 F.2d 812, 822 (5th Cir. 1980) (use of a test for ranking "is justified only if there is evidence showing that those with a higher test score do better on the job than those with a lower test score"); see also Bradley, 443 F.Supp.2d at 159 ("[W]hen an examination is a ranking mechanism that dictates whether and when passing candidates are reached for consideration, the Court must determine whether it is a gateway that has a disparate impact on minority hiring."); 29 C.F.R. § 1607.14(C)(9) (a heightened burden of proof is required when a selection device is used to make employment decisions in strict rank order, requiring a defendant to demonstrate a relationship between higher scores and better job performance).

The Commonwealth claims that the 2005, 2006, 2007, 2008, 2010, and 2012 statewide exams, as well as the 2005 and 2008 Boston exams, were job related because they tested a sufficient number of KSAs "which [could] be practically and reliably measured and which [were] actually required to perform the primary or dominant duties of the position" of sergeant. G.L. c. 31, § 16. The court disagrees. The tests were not sufficiently job related to justify their adverse impact on Blacks and Hispanics.

HRD's examination largely tested candidates' ability to memorize technical knowledge through the use of questions that were often "definitional" or otherwise asked candidates to identify various theoretical concepts with little practical utility. The exams were not representative of the job of a police sergeant, because there was no attempt to test for many critical skills and abilities of police sergeants identified either by HRD in its 1991 Validation

Report, the 2000 Morris & McDaniel Report or numerous reports of experts and segments by HRD managers. The 1991 Validation Report— implausibly and without explanation— states that HRD's examinations could measure skills and abilities that, upon closer review (or even modest scrutiny) could not logically or possibly be tested by a multiple-choice examination or E&E component. HRD failed to meet professional standards in creating its examinations, with the result that, to a significant extent, it assessed test-taking skills, rote memory or theoretical knowledge and absence of test-related anxiety, rather than practical knowledge and critical KSAs. This explains why a large portion of incumbent sergeants failed HRD's examinations when they took a subsequent statewide promotional examination that overlapped with questions on the sergeants' examination. The E&E component does not materially contribute to the validity of HRD's format, because the E&E criteria credit only a limited number of attributes, omit many significant types of prior experience and received nominal weight in a candidates' total score. There is no credible evidence to justify the use of strict rank order as a selection device. Many experts, including HRD's own experts and test plans, indicated that wide score bands were appropriate because HRD's exams were unreliable when used for selection based upon differences in test scores as small as a single point.

It follows that HRD's exams were not reasonably representative of police supervisory duties and were not valid as a device for selecting sergeants.

III. HRD's Knowledge of Disparate Impact and Lack of Job Relatedness

Third, the plaintiff must establish that the defendant knew that the challenged employment practice (1) had a disparate impact on a protected class or group and (2) was not reasonably related to job performance. Lopez II, 463 Mass. at 711.

a. Knowledge of Disparate Impact

HRD knew its 2005, 2006, 2006, 2008, 2012 statewide exams, and the 2005 and 2008 Boston exams "had a significant disparate impact on" Black or Hispanic candidates seeking to be promoted to sergeant in their respective police departments. See Lopez II, 463 Mass. at 711. For instance, HRD's own experts (including Dr. Landy in depositions in 1996 and 2001) have testified that its promotional examinations have an adverse impact on minority candidates. Throughout the 1990s and early 2000s, testimony and reports by other experts, including Dr. Daum and Dr. Lundquist, demonstrated the adverse impact of HRD's examinations on minority candidates. HRD received these reports and testimony in the litigation regarding the disputed examinations.

HRD also created or commissioned analyses showing adverse impact of its examinations. Those reports include the 1991 Validation Report, which found adverse impact on the 1991 statewide examinations at various passing scores. Likewise, a 2006 internal report found severe and statistically significant adverse impact on the 2005 Boston examination. Indeed, rather than present expert testimony to disprove the exams' likely adverse impact, HRD presented Dr. Silva, who challenged Dr. Wiesen's broad conclusions, but found statistically significant adverse impact within numerous multiple departments even without aggregating statewide data for a statewide exam.

b. Knowledge of Lack of Job Relatedness

HRD used the 2005, 2006, 2007, 2008, 2010, and 2012 statewide exams, and the 2005 and 2008 Boston exams, "knowing that [they] were not reasonably related to [the] job performance" of a sergeant. See Lopez II, 463 Mass. at 711. HRD knew of the Civil Service Commission's holding in Carr and the Massachusetts Appeals Court's subsequent endorsement

thereof. See Joint Ex. 38 at 18-19 ("When all the voluminous evidence is brought to bear upon the issue of examination validity, the Commission concludes that the final configuration of the lieutenant's examination – containing only the multiple choice and training and experience components – failed to test for supervisory ability and therefore was not a fair test of the applicants' ability to perform the primary or dominant skills of the position [as required under M.G.L. c. 31 § 16]"); Boston Police Superior Officers Fed'n v. Civil Service Comm'n, 35 Mass. App. Ct. 688, 695 (1993) ("[t]he commission properly found that the multiple choice and training and experience components alone failed to constitute a fair test of supervisory skills and ability.").

Moreover, HRD's 1991 Validation Report identifies a significant number of KSAs that are critical to do the job of police sergeant but could not be tested on a multiple-choice examination. Indeed, the 1991 Validation Report noted that "[t]he assessment of the performance of these skills and abilities would require the use of selection devices outside the scope of the written, multiple-choice format." HRD was aware of the similar statements in the Civil Service Commission and Appeals Court cases where HRD was a party:

> There was substantial evidence before the commission to support its finding that the administrator committed error in deciding that the examination was fair without the performance component. Obviously, supervisory skills and abilities represent a significant element of fitness to perform the primary duties of a Boston police lieutenant. The commission heard expert testimony on the matter. . . . The commission properly found that the multiple choice and training and experience components alone failed to constitute a fair test of supervisory skills and ability.

Boston Police Superior Officers Fed'n, 35 Mass. App. Ct. at 694-695 (affirming the finding that multiple choice and E&E exam alone was not a valid test for a supervisory role "[i]n view of the Legislature's goal that the promotional examinations fairly test the applicants' fitness to perform the primary or dominant duties of the position sought."). Nearly 20 years later, this court reaches

the same conclusion independently, and on the basis of an entirely new record. At the risk of understatement, it is frustrating that HRD has to learn this lesson yet again.

In 2000, Morris & McDaniel also stated that a number of important KSAs could not be tested appropriately in a written exam and recommended that HRD adopt non-written examination components, such as an assessment center or performance review system, and weigh them as heavily as the multiple-choice examination. While HRD claimed that a number of other skills could be tested on either the multiple-choice examination or E&E component, the assertion was implausible on its face. Other experts, including Drs. Daum and Lundquist in the 1990s and Dr. Jacobs in the 2000s, recommended that HRD band candidates' scores because the multiple-choice examination is not a reliable measure of job performance. HRD personnel referred to banding as a best practice. Numerous local police departments, including the City of Boston and the MBTA, have taken the position formally and in court that HRD's promotional examination was not sufficiently valid to justify rank order selection, and that it was thus appropriate to hire out-of-order by selecting minorities who had nominally lower scores.

IV. Knowledge of Less Discriminatory Alternatives

Finally, as in the Title VII disparate impact framework, even if a defendant meets its burden of demonstrating validity, the plaintiff can still prevail if they show that HRD knew there was "another selection device without a similar discriminatory effect that would also serve the employer's legitimate interest." See Bradley, 443 F. Supp.2d at 156, 174 ("Even if the HRD had properly validated the written cognitive examination for use as the sole basis for rank ordering, the plaintiffs have demonstrated the availability of alternative selection devices with less discriminatory effects that would validly serve the HRD's legitimate interests."). "The proper inquiries in the disparate impact analysis are whether the challenged actions were job-related and

consistent with business necessity, and, if so, whether the employer has refused to adopt an alternative employment practice that has less disparate impact and serves the employer's legitimate needs." Abril-Rivera v. Johnson, 806 F.3d 599, 608 n.9 (1st Cir. 2015).

HRD "refuse[d] to adopt an available alternative [sergeant's exam] that has less disparate impact and serve[d] [HRD's] legitimate needs." See Ricci v. DeStefano, 557 U.S. 557, 578 (2009) (citing 42 U.S.C. §§ 2000e–2(k)(1)(A)(ii) and (C)); Compare Lopez I, 823 F.3d at 120 (holding that plaintiffs failed to offer any evidence that the use of assessment center components would have led to a smaller disparity in outcomes, especially given the selection ratios for sergeant promotions in Boston).

Many less discriminatory alternatives were available to HRD, both within and outside the format of a written exam and E&E component. HRD could have reduced the cognitive load by using fewer written questions, assigning a shorter and more relevant reading list and writing the questions and distractors plainly in a manner that addressed practical, rather than abstract, knowledge and rote memorization. It could have avoided questions drawn largely verbatim from textbooks, because such questions have an unnecessary cognitive load and result in known adverse impact. HRD could have graded the written exam on a pass-fail basis or at least reduced the weight given to the written exam. It could have banded scores within a similar range and graded all candidates within that band as equals, particularly where all experts agreed with the concept of score banding. The banding of scores, as Dr. Jacobs' recommended long ago, would allow HRD to consider "long and important job performance records upon which they can be judged." HRD's present 80-20 examination fails to do that.

At the time at issue, HRD was also aware of the use of assessment center techniques, and in fact had approved their use in other municipalities. Dr. Wiesen, Dr. Rafilson, and Dr. Silva all

testified regarding the availability of assessment center alternatives, including in jurisdictions as large as the City of Chicago, where applicants recorded their responses to situational prompts and submitted those responses to be graded. Those alternatives were available and in use in the 2000s. HRD also could have adopted a career board and amended the E&E component to consider job-relevant characteristics that its format ignored. It could have implemented situational judgment exercises in written and non-written forms.

All of these alternatives would have had less impact and would actually have improved the ability to identify the best candidates for promotion to sergeant.

V. Public Interest

The Commonwealth suggests that there is an additional step: assessment of the "public interest." The above analysis under § 4(4A), however, already incorporates the public interest, as defined by the Legislature.

G.L. c. 31, § 16 provides that HRD's "[e]xaminations shall fairly test the knowledge, skills and abilities which can be practically and reliably measured and which are actually required to perform the primary or dominant duties of the position for which the examination is held." That overlaps the court's analysis under G.L. c. 151B, § 4(4A) and is fully consistent with c. 151B One cannot "fairly test" KSAs through a biased test that is not validated to ensure substantial job-relatedness. HRD has not explained how it could violate § 4(4A) and still meet the "fairly test" requirement. Even if that were possible, HRD did not even comply with § 16, considered by itself, because it failed to test many KSAs that "can be practically and reliably measured" and tested for many skills and abilities (such as test-taking and memorization skills) that are not "actually required to perform the primary or dominant duties of the position" of police sergeant.

In any event, the Commonwealth's civil service law and anti-discrimination statute act together to prohibit discrimination in public employment. G.L. c. 151B, § 4(4A) unequivocally sets forth the public interest against interference with the plaintiffs' right to be free from racial and national origin discrimination in promotion to police sergeant. The Civil Service law explicitly incorporates the same fundamental public policy. The civil service law expressly mandates that decisions be consistent with "basic merit principles." Massachusetts Ass'n of Minority Law Enforcement Officers v. Abban, 434 Mass. 256, 264 (2001) (fundamental purpose of civil service law is "to ensure decision-making in accordance with basic merit principles"). In defining "basic merit principles, G.L. c. 31, § 1 provides: "'Basic merit principles' [include] "assuring fair treatment of all applicants and employees in all aspects of personnel administration without regard to political affiliation, race, color, age, national origin, sex, marital status, handicap, or religion and with proper regard for privacy, basic rights outlined in [G. L. c. 31] and constitutional rights as citizens." See also Brookline v. Alston, 487 Mass. 278, 294-297 (2021) (harmonizing c. 31 civil service laws with anti-discrimination laws under G.L. c. 151B and c. 31 in the context of a racially tainted motivation for termination of employment); Boston Police Superior Officers Fed'n, 35 Mass. App. Ct. at 695.

HRD apparently argues that the civil service law ties its hands. But nothing in this decision undermines the requirements of G.L. c. 31, § 25, which provides that: "[t]he names of such persons [candidates] shall be arranged on each such list . . . in the order of their marks on the examination based upon which the list is established." For one thing, an "examination" may be oral, written or a combination of the two. It may occur in a multiple choice test or in other standard assessment methodologies recognized by E.B. Jacobs and others in the facts found above. Moreover "marks" need not be specific numbers scored on a multiple choice exam.

They could be the equivalent of an academic letter grade, which bands groups of numerical scores into a single grade. Finally, if one takes a very literal approach, as HRD urges, even the educational and experience component might stretch the concept of "marks on the examination," because the award of points for specific aspects of experience or education is entirely judgmental, without scientific or empirical basis. That narrow reading of HRD's authority is not consistent with statutory language on purpose.[10]

It follows that nothing in § 25 conflicts with developing a promotional system that avoids interference with plaintiffs' rights to be free of interference from racial or national origin discrimination in promotion to sergeant. Nor does it conflict with the notion that "competitive exams [are] a tool to accomplish an important public policy of moving away from nepotism, patronage, and racism in the hiring and promoting of police." Lopez I, 823 F.3d at 108. On the contrary: a competitive exam (written, oral or both) that avoids disparate impact and reasonably reflects actual job qualifications will do a far better job of moving away from racism in promotions.

VI. MCAD FILINGS

The court ruled before trial that the plaintiffs had exhausted their remedies by filing a timely complaint with the Massachusetts Commission against Discrimination. See Everett v. 357 Corp., 453 Mass. 585, 600 (2009) ("Without the predicate filing in MCAD, the Superior Court has no jurisdiction to entertain the claim of discrimination."). See also Lewis v. City of

[10] HRD has legal authority to implement banding, because, as noted above, a band still establishes a rank order based upon scoring, much as a letter grade system does (e.g. by treating scores from 93 to 97 as an "A", without differentiation). Indeed, HRD's actual practice of rounding to the nearest whole number constitutes banding of all scores within 0.50 points of that whole number. To the extent that HRD relies upon the preliminary injunction in Pratt v. Dietl, Suffolk Superior Court No. 09-1254 (April 16, 2009), that decision was not only preliminary, but also turned upon (at p. 10) the failure to follow "the requisite review process called for by the statute." To remedy the violation in this case, therefore, HRD only had to follow the statutory process to implement banding.

Chicago Ill., 560 U.S. 205, 212 (2010) (in a Title VII disparate impact claim against an employer, the time to file an EEOC charge begins to run each time the employer makes a selection from an eligible list). In denying the "Defendants' Motion to Dismiss Plaintiffs' Claims Related to the 2005, 2010 and 2012 Exams for Lack of Subject Matter Jurisdiction," the court made the following filings and rulings, which it readopts and reiterates:

> The Commonwealth seeks dismissal of all claims related to the 2005 Exam, 2010 Exam and 2012 Exam on the ground that no named plaintiff or class member filed an MCAD charge related to these exams within the 300-day statute of limitations. See G.L. c. 151B, § 5. The named plaintiffs' MCAD charges, filed in December, 2007 through September, 2008, alleged that the earliest date on which discrimination occurred was in 2007 or 2008. The 2005 exam was more than 300 days [before] those dates.
>
> It is true that, "[w]ithout the predicate filing in MCAD, the Superior Court has no jurisdiction to entertain the claim of discrimination." Everett v. 357 Corp., 453 Mass. 585, [600] (2009). However, for jurisdictional purposes in this class action, there is a predicate filing.
>
> The 300-day period begins to run when the plaintiff knew or should have known of the alleged discriminatory act. Flint v. City of Boston, 94 Mass. App. Ct. 298, 303[] (2018). The Commonwealth claims that the period starts running "when the eligibility lists were issued for each exam . . . ." Comm. Mem. at 5. This is incorrect. Lewis v. City of Chicago, Ill., 560 U.S. 205, 130 S.Ct. [2191,] 2199 (2010) (Rejecting a statute of limitations argument based upon promulgation of a promotional list based upon an invalid test, because "[i]f petitioners could prove that the City "use[d]" the "practice" that "causes a disparate impact," they could prevail."). Here, the plaintiffs offer to prove that promotions were being made from the 2005 exam list in late 2007 and throughout 2008. Where, as here, there is no requirement to show intent, the continuing impact is enough. Id. The court recognizes, as the Commonwealth's Reply points out, that the defendant in Lewis was the appointing authority, not the entity that provided tests, scored them and ranked candidates based upon test result. Where the HRD's list was in effect over a period of time, HRD had the authority to take necessary action to avoid disparate impact during that time, failed to do so, and thereby permitted the alleged disparate impacts to continue and recur, that distinction does not make a difference in a disparate impact case. It follows that the court has jurisdiction to adjudicate those claims.
>
> Moreover, where the MCAD complaints included broad language about disparate impacts on minority test takers and lack of job relatedness, the "scope of the complaint" rule applies to the 2010 and 2012 exams, as well as statewide, to all municipalities affected by the HRD practices in question. Id. See Pelletier v. Town of Somerset, 458 Mass. 504, 514 (2010) ("the MCAD [charge] and potential investigation establish the scope of any subsequent filing in the Superior Court."). Where the plaintiffs challenged a

set of consistent practices that applied to all exams, lists and scoring at issue, the scope of the MCAD's investigation reasonably would include the subsequent, substantially identical practices, challenged in this case. See Everett, 453 Mass. at 603. All that is required for jurisdictional purposes, is that the conduct in the underlying MCAD complaint is "reasonably related" to the claims in this case. Id.

Finally, the "single filer" or "piggybacking" rule likely applies to this disparate impact class action discrimination case. See, e.g. Perez-Abreu v. Metropol Hato Rey LLC, 5 F.4th 89, 92 (1st Cir. 2021). Under this rule, the administrative filing of one class member satisfies the filing obligations of all class members. Id. Here, there are 73 administrative charges, of which at least 45 are not precluded by prior litigation. While Massachusetts has not explicitly adopted the single filer rule, the court believes that the Supreme Judicial Court is likely to adopt it, because of the solid federal precedent on that rule and because the alternative would be the very proliferation of claims and complexity that the class action device is designed to avoid.

The court makes the following additional findings:

209. 53 individuals filed MCAD charges related to the exams at issue in this case. The evidence includes MCAD charges filed by 44 individuals. It does not include MCAD charges filed by anyone who took the 2010 or 2012 statewide sergeant's exam, but the scope of the complaint rule meets the filing requirement for those exams.

210. Named Plaintiff Spencer Tatum took only the 2006 and 2008 statewide exams.

211. Named Plaintiff Louis Rosario took only the 2005 and 2007 statewide exams.

212. Named Plaintiff Francisco Baez took only the 2006, 2008, 2010, and 2012 statewide exams.

213. The following Boston police officers were named plaintiffs in Lopez v. City of Lawrence et al., Case No. 1:07-cv-11693-GAO (D. Mass. 2014): Gwendolyn Brown, Shumeane Benford, Angela Williams-Mitchell, Lynette Praileau, Tyrone Smith, Eddy Chrispin, David E. Melvin, Steven Morgan, William E. Iraola, Jose Lozano, Courtney A. Powell, James L. Brown, George Cardoza, Larry Ellison, David Singletary, Charisse Brittle-Powell, Cathenia D. Cooper-Paterson, Molwyn A. Shaw, Lamont Anderson, Gloria Kinkead, Kenneth Gaines, Murphy

Gregory, Julian Turner, Neva Grice, Delores E. Facey, Lisa Venus, Rodney O. Best, Karen VanDyke, and Robert C. Young.

214. 29 of the 44 MCAD charges offered into evidence by Plaintiffs were filed by Boston police officers who were named plaintiffs in Lopez I.

215. 7 Boston police officers were not named plaintiffs in Lopez I who filed MCAD charges. They filed their MCAD charges on April 1, 2010.

216. The eligible list for the 2005 Boston exam was published on February 13, 2006, and expired on May 14, 2009.

217. No candidate who took the 2005 Boston exam and who was not a named plaintiff in Lopez I filed an MCAD charge by March 22, 2010.

218. With a few exceptions that are not relevant, the eligible lists for the 2005 statewide exam were published on March 24, 2006 and expired on March 30, 2008.

219. No candidate who took the 2005 statewide exam filed an MCAD charge by January 18, 2007.

220. The class properly includes those who took the 2010 and 2012 exams.[11]

---

[11] The court has rejected Defendants' claim that the Third Amended Class Action Complaint did not include minority candidates for the 2010 and 2012 examinations. The court previously allowed Plaintiffs' motion that explicitly stated that "Plaintiffs seek ... to add the sergeant promotional exams for 2010 and 2012." See Docket No. 38, Plaintiffs' Mot. to Sub. Named Plaintiff and For Leave to File Third Amend. Compl. at 1; see also id. at 2 ("Plaintiffs also seek to add the promotional exams for sergeant for 2010 and 2012. The complaint currently includes the 2005, 2006, 2007, and 2008 promotional exams for sergeant."). In their opposition to this motion, Defendants explicitly acknowledged that Plaintiffs were moving "amend to include tests administered in 2010 and 2012", see Docket No. 38, Defendants' Opp. to Plaintiffs' Third Mot. to Amend Compl. at 1. The defendants did not mention the scrivener's error in paragraph 10 of the third amended complaint. The Court unconditionally allowed Plaintiff's motion to amend on February 4, 2014, see Docket No. 39, and in the ensuing seven-and-a-half years, all parties have litigated this case as if minority candidates for the 2010 and 2012 examinations were included in the class. The Defendants cannot legitimately claim that the class does not encompass those claims. Even if they could, they waived that claim by failure to raise it in a timely manner, so that it could be addressed and cured.

CONCLUSION

Overwhelmingly persuasive evidence proves that HRD interfered with the class members' rights to consideration for promotion to police sergeant without regard to race or national origin. HRD failed to implement some very simple ways to reduce adverse impact upon Black and Hispanic candidates. Artificial reduction of the eligible pool resulted in consideration of fewer candidates overall, including minority candidates. That, in turn, precluded considering many candidates on their full merits, as opposed to their test scores.

Instead of improving its assessment format, HRD promulgated lists to provide a thin veneer of apparent justification for a discriminatory process. The false appearance of a fair process created inaccurate beliefs and created unwarranted expectations among candidates and appointing authorities. Those beliefs and expectations have had a life of their own in perpetuating a discriminatory system that has injured qualified candidates and deprived the public of the benefits of having the best-qualified police sergeants. In all these actions, HRD knew what it was doing.

On these facts, HRD most certainly violated § 4(4A) with respect to the plaintiff class for all the exams at issue.

Dated: October 27, 2022

/s/Douglas H. Wilkins
Douglas H. Wilkins,
Justice of the Superior Court