443 F.Supp.2d 145 (2006)

Jacob BRADLEY, Noah Bradley, Keith Ridley, and Jared Thomas, Plaintiffs,

v.

CITY OF LYNN, et al., Defendants.

Civil Action No. 05-10213-PBS.

United States District Court, D. Massachusetts.

August 8, 2006.

148 *146 *147 *148 Nadine M. Cohen, Lawyers' Committee for Civil Rights Under Law, Mark D. Selwyn, Wilmer Cutler Pickering Hale and Dorr LLP, Alfred Gordon, Harold L. Lichten, Shannon E. Liss-Riordan, Pyle, Rome, Lichten, Ehrenberg & Liss-Riordan, P.C., Boston, MA, for Plaintiffs.

George S. Markopoulos, James Lamanna, City Solicitor's Office, Lynn, MA, Ronald F. Kehoe, Sookyoung Shin, Attorney General's Office, Boston, MA, for Defendants.

MEMORANDUM AND ORDER

SARIS, District Judge.

I. *INTRODUCTION*

In this class action, the plaintiffs[1] allege that the written civil service cognitive ability examination used in 2002 and 2004 to qualify and rank applicants has had a disparate and adverse impact on Black and Hispanic candidates for entry-level firefighter positions in violation of Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e-2(a), (k) (2006), and the federal consent decree in *Boston Chapter, NAACP, Inc. v. Beecher,* 371 F.Supp. 507 (D.Mass.1974) (the *"Beecher* decree"). The defendants are the Human Resources Division of the Commonwealth of Massachusetts (the "HRD"), which develops and administers the examination, the City of Lynn, and various public officials. The HRD argues that because of the statutory veterans preference, residency requirements, and other selection factors, the examination has no disparate impact on the bottom-line hiring of Black and Hispanic candidates for entry-level firefighter positions in Massachusetts.

Pursuant to Fed.R.Civ.P. 23(a) and (b)(2), the Court certified the plaintiff firefighter and police officer classes on March 24, 2006. The Court certified the firefighter class as "[a]ll minorities (Black and Hispanic) who took the civil service examination for the position of fire fighter within the Commonwealth of Massachusetts in the years 2002 and 2004." (Docket No. 81.) A six-day bench trial for the liability phase of the firefighter class began on April 11, 2006, and the parties rested on May 4, 2006.

The following witnesses testified for the plaintiffs: Dr. Frank Landy, an expert in industrial psychology and
149 statistics and a former consultant to the HRD; Dr. Joel *149 Wiesen, an expert in industrial psychology and statistics and the HRD Chief of Test Development and Validation between 1977 and 1993; Elizabeth Dennis, the former HRD Director of the Civil Service Unit between the mid-1990s and 2003; and Kevin Bradley, a Lynn firefighter since 1977 and the father of two of the named plaintiffs. The following witnesses testified for the HRD: Dr. Rick Jacobs, an expert in industrial psychology and statistics and a former and current consultant to the HRD; Marc Chavanne, the HRD Deputy Director of Selection and Validation between 1991 and 1996; and Sally McNeely, the current, HRD Director of the Organizational Development Group since 2003.

The parties submitted closing briefs, and oral argument was held on June 9, 2006. After trial, oral argument, and review of the post-trial submissions, the Court holds that the written civil service cognitive ability examinations used in 2002 and 2004 have an adverse and disparate impact on the employment opportunities of Black and

Hispanic candidates for entrylevel firefighter positions, and that the selection process that uses the examination scores to rank candidates is not job related and consistent with business necessity under applicable federal law and the longstanding *Beecher* decree. The plaintiffs have also demonstrated that there are alternative selection methods with less discriminatory effects that would serve the employer's legitimate interest in selecting capable firefighters based in part on cognitive ability. Accordingly, I conclude that judgment on liability should enter in favor of the plaintiff firefighter class.

II. *FINDINGS OF FACT*

A. *Statutory and Administrative Framework*

The civil service law currently applies to the fire departments of approximately 110 municipalities in Massachusetts, including Boston and Lynn. To become a firefighter in a municipality where the civil service law applies, an individual must first pass a statewide civil service examination. *See* Mass. Gen. Laws ch. 31, §§ 6, 59.

The personnel administrator for the HRD (the "HRD Administrator") conducts, determines the form, method, and subject matter of, and develops the examinations, *id.* §§ 5(e), 16; prepares and posts notices of the examinations, *id.* §§ 18-19; and determines the passing requirements, *id.* § 22. Based on the examination results, the HRD Administrator ranks the names of those who pass on the "eligible list" based on the following statutory priority. *Id.* §§ 25, 26.

> The names of persons who pass examinations for original appointment to any position in the official service shall be placed on eligible lists in the following order: (1) disabled veterans, in the order of their respective standings; (2) veterans, in the order of their respective standings; (3) widows or widowed mothers of veterans who were killed in action or died from a service connected disability incurred in wartime service, in the order of their respective standings; (4) all others, in the order of their respective standings.

Id. § 27.

To hire for a firefighter vacancy, a municipality's appointing authority submits a request to the HRD Administrator, who then certifies "from the eligible list sufficient names of persons for consideration" in rank order. *Id.* § 6. In addition to the statutory priority, the appointing authority may have the HRD Administrator rank residents ahead of
150 non-residents, *see id.* § 58, and may request special certification lists for candidates with certain qualifications,
*150 such as Spanish-language abilities (Ex. 9, at 16 (Pers. Admin. Rule 8(4))). The HRD and the HRD Administrator (the "State defendants") have interpreted the civil service law as giving them discretion to decide how many names should be certified from the eligible list.

Once the HRD Administrator certifies a list to a municipality, each candidate must sign the certified list and express a willingness to accept employment in order to be considered for appointment. Mass. Gen. Laws ch. 31, § 25. According to the HRD's Personnel Administration Rules, the appointing authorities for a municipality "may appoint only from among the first 2n + 1 persons named in the certification willing to accept appointment," where "n" is the number of vacancies. (Ex. 9, at 16-17 (Pers. Admin. Rule 9(1)).) For five vacancies, for example, a municipality may only appoint from the first eleven named in the certification willing to accept. *(See id.; see also* Exs. 33P, 33Q (stating on certification lists that selection "must be" within first 2n + 1 "who will accept").)

In evaluating the candidates within the "2n + 1" pool, a municipality may establish its own hiring criteria, such as a drug test, a background check, or an interview. After conditional offers are made by the municipality, the HRD administers a pass/fail physical abilities test. Some municipalities conduct a full medical or psychological examination. According to Ms. McNeely, some municipalities use these post-certification hurdles to determine the composition of the "2n + 1" pool.

None of the post-certification hurdles, however, changes a candidate's ranking on the list. By statute, to bypass higher-ranked individuals on the certified list to hire lower-ranked individuals, a municipality must submit a written statement to the HRD Administrator justifying the bypass. Mass Gen. Laws ch. 31, § 27. The HRD Administrator

has the right to review and withdraw any bypass appointment. *MacHenry v. Civil Serv. Comm'n,* 40 Mass.App.Ct. 632, 634-36, 666 N.E.2d 1029, 1030-31 (Mass.App.Ct.1996).[2]

B. *Beecher Firefighter Litigation*

The civil service examination for firefighters has been the subject of employment discrimination litigation since the 1970's. The *Beecher* class action was brought by the Boston Chapter, NAACP, on behalf of a statewide class of Black and Spanish-surnamed applicants for the firefighter position. *Beecher,* 371 F.Supp. at 509-10. In *Beecher,* although the available examination statistics were "meager," after comparing the minority population and employment statistics, the district court concluded that the plaintiffs established a prima facie case that the written examination had a discriminatory effect on Blacks and Spanish-surnamed persons and that the defendants failed to demonstrate that the examination was substantially related to job performance. 371 F.Supp. at 514, 517.

As a result, the district court issued a consent decree, which established certification quotas for minorities in "all cities and towns subject to Civil Service law" until "a city or town achieves a complement of [firefighter] minorities commensurate with the percentage of minorities within the community." *Id.* at 522-23. Importantly, the decree ordered:

> The Massachusetts Division of Civil Service shall cease using written firefighter entrance examinations of the type administered by the Division of Civil Service in August 1971, for the
> 151 purpose of *151 determining qualifications for the selection of firefighters. Should the Division of Civil Service desire to utilize entrance examinations in the future for the purpose of selecting firefighters, such examinations shall be demonstrably job-related and validated in accordance with the "Guidelines on Employees Selection Procedures" issued by the Equal Employment Opportunity Commission, 29 C.F.R. § 1607.1 et seq., or otherwise shown to have no discriminatory impact. If the parties disagree as to whether a written examination has been shown to be valid within the meaning of the Guidelines,,the question of their validity and job relatedness shall be resolved by the Court, and such resolution, whether by the parties' agreement or by the Court, shall be accomplished before any such test is put into use for the purpose of qualifying or selecting. As with the instant study, the Court will scrutinize closely a future study which shows only a minimal level of job-relatedness.

Id. at 521.

The First Circuit affirmed the district court's disparate impact findings in *Boston Chapter, NAACP, Inc. v. Beecher,* 504 F.2d 1017, 1021, 1026 (1st Cir.1974). With respect to the prima facie burden, the First Circuit stated: "Plaintiffs usually meet their initial burden by demonstrating that minority candidates have a higher test failure rate." *Id.* at 1019. In addition, the First Circuit found that the validation study did not survive close scrutiny and affirmed the *Beecher* decree's implementation of certification quotas that remained in effect for each local fire department "until that department attains sufficient minority fire fighters to have a percentage on the force approximately equal to the percentage of minorities in the locality." *Id.* at 1024-28.

Over the past thirty-plus years, municipalities have been released from the *Beecher* decree as their fire departments achieved racial parity with their populations. As of March 14, 2006, only nine of the 110 municipalities subject to the civil service law remain under the decree. (Ex. 8.) Since the *Beecher* decree ended in Lynn in 1986, Mr. Bradley has estimated that only four of the 106 entry-level firefighters hired in Lynn have been Black or Hispanic. Since the First Circuit held that the *Beecher* decree should no longer apply to Boston in *Quinn v. City of Boston,* 325 F.3d 18, 37 (1st Cir.2003), as of the time of trial, only seven of the 105 entry-level firefighters hired in Boston from the 2004 examination have been reported to be Black or Hispanic. (*See* Ex. 33D, at 2.)

C. *Firefighter Hiring from the 2002 and 2004 Civil Service Examinations*

The HRD administered a civil service examination for firefighters on April 27, 2002 (Ex. 28) and another on April 24, 2004 (Ex. 30). Both examinations contained one-hundred multiple choice questions testing only cognitive

ability.[3] The 2002 examination tested 4543 applicants, and the 2004 examination tested 2447 applicants. (*See* Ex. 5, Table 1; Ex. 6, Table 1.)

152 *152 The HRD used the passing point of seventy, an arbitrary number that has been used since at least 1971. *See, e.g., Beecher,* 371 F.Supp. at 511-12. The HRD adjusted scores after administering the examination both by removing questions and by crediting multiple answers as correct on questions so that the passing point of seventy produced no adverse impact on minorities under the EEOC Guidelines. (Ex. 10, at 3; Ex. 12, at 4-5; Trial Tr. 23, May 3, 2006; Trial Tr. 144-46, May 4, 2006.)

Before putting the 2002 and 2004 examination results into use for hiring, the HRD provided adverse impact analyses to the court-appointed NAACP monitor for the *Beecher* decree. (Exs. 10, 12.) The analyses for both examinations showed no adverse impact at the passing point of seventy under the EEOC Guidelines; however, the analyses did show that minorities were adversely impacted at every score above seventy. (Ex. 10, Attach. I; Ex. 12, Attach. G.) While stating to the HRD that the "examination results reflect a significant adverse impact on Black and Hispanic candidates at scores higher than 70," the NAACP monitor did not object before the HRD put the 2002 and 2004 examination results into use for firefighter hiring. (Ex. 13.)

Based on the certified lists the HRD Administrator provided in response to requests from municipalities, 311 candidates were hired from the 2002 examination, and thus far, 200 candidates have been hired from the 2004 examination. (*See* Ex. 5, Table 1; Ex. 6, Table 1.) The overall hiring numbers in Massachusetts indicate that minority candidates have been hired less frequently than non-minority candidates:

Civil Service Exam	2002	2004
Number of Minority Takers	555	502
Number Minorities Appointed	19	16
Minority Appointment Rate	3.4%	3.2%
Number of Non-Minority Takers	3988	1945
Number of Non-Minorities Appointed	292	194
Non-Minority Appointment Rate	7.3%	10%
Ratio of Appointment Rates Between Minorities and Non-Minorities	47%	32%

(*See* Ex. 5, Table 1; Ex. 6, Table 1.)

There will be further hiring from the pool of candidates who passed the 2004 examination. For example, Boston, which hired a class in June 2005 and January 2006, recently requested a new list of candidates based on the 2004 examination to hire for fifty vacancies. The HRD certified a list of candidates to Boston on April 11, 2006, and expanded it on April 21, 2006, providing a total of 156 candidates. The expanded list was produced to the plaintiffs on May 4, 2006, the last day of testimony. (Exs. 33M, 33Q.) Also on May 4, 2006, the HRD certified thirty-seven candidates in response to Lynn's request to fill four vacancies. (Ex. 33P.)

D. *Creation of the Entry-Level Civil Service Examination for the Rank Ordering of Candidates*

Beginning in 1992 and continuing throughout the 1990's, the HRD hired Landy, Jacobs and Associates, Inc. ("Landy-Jacobs") to develop the written cognitive examinations. Dr. Landy is now the plaintiff's expert; Dr. Jacobs is the HRD's expert. While no longer affiliated, they remain good friends.

In June 1992, Landy-Jacobs completed the Massachusetts Firefighter Final Validation Report (the "1992 Report"). (Ex. 27.) The 1992 Report was done under the direction of Dr. Landy and concluded that the written
153 and physical examinations proposed by Landy-Jacobs for use in the *153 selection of firefighters in Massachusetts were valid. Criterion-related validity evidence[4] existed for both the written and physical examinations from a study performed in 1986 for the Columbus, Ohio Fire Department. (*Id.* at 7.) The 1992 Report also documented a "high degree of demonstrated similarity between the job of firefighter in Massachusetts and the job of firefighter in Columbus," indicating the transportability of the written and physical examinations developed in Columbus, Ohio to Massachusetts. (*Id.* at 1-2.) In addition, the 1992 Report found that the written and physical examinations could validly be used for rank-order selection. (*Id.* at 5.) Importantly, Landy-Jacobs did not validate tile written cognitive examination for rank ordering as a stand-alone mechanism; rather, the 1992 Report validated rank ordering only when the written examination constituted 40% and the physical examination constituted 60% of the overall composite score. (*See* Trial Tr. 5-9, Apr. 13, 2006 ("[W]hat we did was to validate the procedure, which was a combination of a weighted cognitive ability and a weighted physical ability exam."); Trial Tr. 8-9, May 4, 2006 ("[T] he weighting [Landy-Jacobs] recommended should be 60 percent physical and 40 percent cognitive.").)

Criterion-related validity studies, such as the 1986 Columbus, Ohio study, determine whether "the selection procedure is predictive of or significantly correlated with important elements of job performance." 29 C.F.R. § 1607.5(B). The magnitude of the relationship is measured by calculating a correlation coefficient. *Id.* § 1607.14(B)(6).

In the 1986 Columbus, Ohio study, Landy-Jacobs found the correlation coefficient of the cognitive ability exam to be between 0.2 and 0.3 (Ex. 24, Tab 6, at 2; Ex. 27, Attach. A.) The physical abilities test was found to have a correlation coefficient between 0.3 and 0.4. (Ex. 27, Attach. 9.) These correlation coefficients were one of the reasons why Landy-Jacobs believed that the physical agility test should comprise 60% of the total score. Landy-Jacobs never validated the cognitive examination to be the exclusive basis for rank ordering.

The HRD purchased the written cognitive examination from Landy-Jacobs, administered it on May 22, 1993, and weighted it at 40% of the overall score. (Ex. 24, Attach. 6, at 1, 9.) The other 60% of the overall score came from the physical examination, which the HRD developed, with the help of Landy-Jacobs.

E. *Physical Examination*

The physical examination, which was paired with the 1993 written civil service examination, consisted of several timed events, such as a stair climb event that required candidates to make six trips up and down two flights of stairs carrying different pieces of equipment, a ladder event where a pulley mechanism replicated raising a ladder, and a rescue event that replicated crawling into a dark area to save a victim. (Trial Tr. 8-9, 14-16, May 4, 2006.)

During this selection process, two factors delayed the HRD's ability to issue certified lists for municipal vacancies. First, by placing the physical examination at the beginning of the hiring process as an initial screen rather than at
154 the end as a final pass/fail hurdle, the number of applicants tested physically increased to almost *154 5000. This increase required the HRD to secure and construct additional testing sites, which it did. Second, the HRD had to suspend the administration of the physical examination after receiving medical complaints from eight candidates. An expert panel convened by the HRD concluded that the physical examination could resume with additional safeguards, including making water more accessible and giving more discretion to on-site emergency medical technicians to screen out candidates based on high blood pressure or heart rate. When the HRD resumed administering the physical examination, no further medical problems were reported.

The average physical examination score of non-minority candidates was 95.55 and of minority candidates was 92.43. While 78.6% of non-minority candidates achieved one of the top two possible scores of 100 or 95, only

61.3% of minority candidates scored as well. (Ex. 35B.) Based on these results, the HRD concluded that administering the physical examination at the beginning of the hiring process did not "mitigate the adverse impact [of the written cognitive examination] as much as [the HRD] had hoped it would." The HRD thus "saw no advantage in continuing this beyond 1994" and instead, decided to test for physical abilities at the end when making conditional job offers, enabling full-fledged medical evaluations and avoiding administrative time and costs. (Trial Tr. 68, May 3, 2006; Trial Tr. 24-27, May 4, 2006; Ex. 24, Attach. 6, at 9.) Dr. Landy testified, however, that the difference in physical examination scores was not statistically significant. As such, the use of the physical examination at the beginning of the hiring process, which decreased the weight of the written examination to 40% of the overall score, had the effect of diluting the adverse impact of the written examination. (Trial Tr. 44-47, Apr. 13, 2006.)

Regardless, after 1993, the HRD ultimately decided to use only the written cognitive examination for rank ordering and to administer the physical examination at the end of the hiring process as one of the final pass/fail hurdles for entry-level firefighter candidates who already received conditional offers.

F. *Subsequent Job Analyses and Validation Studies*

In December 1995, Landy-Jacobs completed the Massachusetts Fire Departments Job Analysis and Physical Fitness Standards Test Development Report (the "1995 Report"). (Ex. 36.) The 1995 Report documented a job analysis and validation of physical fitness and medical standards for firefighters of all ranks. (*Id.* at 1.) The job analysis consisted of (a) developing a preliminary list of tasks by using job analyses conducted of fire departments in other cities; and (b) surveying a sample of Massachusetts firefighters to determine whether each task was performed by them, whether a firefighter would be responsible for performing the task on their first day, and the task's importance and frequency. (*Id.* at 4-5.) The 1995 Report concluded that the physical fitness standards test was content-valid.[5] (*Id.* at 11-12.) The development of the physical fitness test included using a sample of incumbent firefighters to set reasonable cut times for each event. (*Id.* at 9.)

In June 2002, SHL USA, Inc. ("SHL"), the firm that purchased Landy-Jacobs, completed the Commonwealth of
155 Massachusetts *155 2002 Entry-Level Firefighter and Police Officer Job Analysis Report (the "2002 Job Analysis"). (Ex. 38.) The 2002 Job Analysis updated and revised the job analysis from the 1996 Report. (*Id.* at 1.) SHL used the findings from the 1996 Report to form a preliminary task list, which SHL enhanced by referencing previously conducted job analyses for entry-level firefighters in other jurisdictions and by interviewing incumbents and supervisors from Massachusetts. (*Id.* at 5-6.) SHL matched the tasks with a set of cognitive and motor abilities based on a published ability taxonomy. (*Id.* at 7-8.) The HRD then administered a SHL-developed survey to determine each task's importance and frequency and each ability's relative importance. (*Id.* at 9-14.) The 2002 Job Analysis concluded that "the basic areas of responsibility for firefighters have changed very little over the years in terms of the importance of the duties relative to one another" and that "the abilities needed to succeed as a firefighter or police officer within the Commonwealth of Massachusetts have not changed substantially over time." (*Id.* at 24.) "This suggests that the existing cognitive and physical examinations should continue to have great relevancy in their current form with only slight enhancements based on the 2002 job analysis results." (*Id.*)

Also in June 2002, SHL completed the Commonwealth of Massachusetts Entry-Level Firefighter and Police Officer Medical Standards Update Report (the "2002 Medical Update"). (Ex. 37.) The 2002 Medical Update documented an update and a content-validation of the then existing medical standards based on the 2002 Job Analysis. (*See id.* at 1-3.)

G. *Creation of the 2002 and 2004 Civil Service Examinations*

The HRD designed the 2002 and 2004 examinations in-house with the goal that they would be equivalent or comparable to the Landy-Jacobs examinations. (Ex. 10, at 1; Ex. 12, at 3.) No outside consultants assisted, and during that time, the HRD employed no industrial psychologists. Dr. Jacobs believes that the 2002 and 2004 examinations are similar to the examination created by Landy-Jacobs for Massachusetts in 1992, and Dr. Landy agrees that the 2002 and 2004 examinations appear to follow the "spirit" of the Landy-Jacobs 1992 examination. Dr. Landy also points out, however, that there is no evidence of the HRD analyzing the 2002 and 2004 examinations to ensure, for example, that the reading level of the new examinations did not exceed the reading level of the job or that the new questions were linked with the constructs intended to be tested. (*See* Ex. 1, at

22-23.) Beginning in June 2006, the HRD administered a new test, which it claims is state-of-the-art and will be validated.

III. *CONCLUSIONS OF LAW*

A. *Title VII*

1. *Statutory Framework*

Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e-2(a), prohibits not only overt and intentional discrimination but also "more subtle forms of discrimination, known as disparate impact discrimination." *EEOC v. Steamship Clerks Union, Local 1066,* 48 F.3d 594, 600-01 (1st Cir.1995). The "disparate impact approach roots out `employment policies that are facially neutral in their treatment of different groups but that in fact fall more harshly on one group than another and cannot be justified by business necessity.'" *Id.* (citations omitted). Stated
156 another way, the disparate impact approach prohibits employment "practices *156 that are fair in form, but discriminatory in operation." *Griggs v. Duke Power Co.,* 401 U.S. 424, 431, 91 S.Ct. 849, 28 L.Ed.2d 158 (1971).

Section 2000e-2(a)(2) provides inrelevant part:

> It shall be an unlawful employment practice for an employer . . .
>
> (2) to limit, segregate, or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities or otherwise adversely affect his status as an employee, because of such individual's race, color, religion, sex, or national origin.

The Civil Rights Act of 1991[6] added a provision to make the burden of proof in disparate impact cases explicit:

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

42 U.S.C. § 2000e-2(k)(1)(A). "The term `demonstrates' means meets the burdens of production and persuasion." *Id.* § 2000e(m). "The demonstration referred to by subparagraph (A)(ii) shall be in accordance with the law as it existed on June 4, 1989, with respect to the concept of `alternative employment practice,'" *id.* § 2000e-2(k)(1)(C), which the courts had defined as another selection device without a similar discriminatory effect that would also serve the employer's legitimate interest. The plaintiff bears the burden of demonstrating the existence of alternative business practices. *Int'l Bhd. of Elec. Workers v. Miss. Power & Light Co.,* 442 F.3d 313, 317-18 (5th Cir.2006).

2. *Shifting Burden*

The First Circuit set forth the legal framework that applies to disparate impact cases in *Steamship Clerks.*

> [I]t is incumbent upon the plaintiff to demonstrate a prima facie case of discrimination. In the disparate impact milieu, the prima facie case consists of three elements: identification, impact, and causation. First, the plaintiff must identify the challenged employment practice or policy, and pinpoint the defendant's use of it. Second, the plaintiff must demonstrate a disparate impact on a group characteristic, such as race, that falls within the protective ambit of Title VII. Third, the

> plaintiff must demonstrate a causal relationship between the identified practice and the disparate impact.
>
> When the plaintiff rests, declaring herself satisfied that she has established a prima facie case of disparate impact discrimination, the ball bounces into the defendant's court. At that point, the defendant has several options. First, it may attack the plaintiff's proof head-on, debunking its
157 sufficiency or attempting to rebut it by adducing countervailing *157 evidence addressed to one or more of the three constituent strands from which the prima facie case is woven, [] asserting, say, that no identifiable policy exists, or that the policy's implementation produces no disparate impact, or that the plaintiff's empirical claims-such as the claim of causation-are insupportable.
>
> Alternatively, the defendant may confess and avoid, acknowledging the legal sufficiency of the prima facie case but endeavoring to show either that the challenged practice is job-related and consistent with business necessity, [] or that it fits within one or more of the explicit statutory exceptions covering bona fide seniority systems, veterans' preferences, and the like. In all events, however, a defendant's good faith is not a defense to a disparate impact claim.
>
> If the defendant fails in its efforts to counter the plaintiff's prima facie case, then the factfinder is entitled-though not necessarily compelled, [] [illegible] to enter judgment for the plaintiff. On the other hand, even if the defendant stalemates the prima facie case by elucidating a legitimate, nondiscriminatory rationale for utilizing the challenged practice, the plaintiff may still prevail if she is able to establish that the professed rationale is pretextual. The plaintiff might demonstrate, for example, that some other practice, without a similarly undesirable side effect, was available and would have served the defendant's legitimate interest equally well. Such an exhibition constitutes competent evidence that the defendant was using the interdicted practice "merely as a `pretext' for discrimination."

48 F.3d at 601-02 (citations and footnotes omitted). While *Steamship Clerks* addressed the legal framework as it existed prior to 1991, the First Circuit has applied the same framework in the context of the Civil Rights Act of 1991. *See Donnelly v. R.I. Bd. of Governors for Higher Educ.,* 110 F.3d 2, 4 (1st Cir.1997).

As part of the prima facie case, the plaintiffs must demonstrate that the civil service examinations have both an adverse and disparate impact. Specifically, the plaintiffs must demonstrate that the adverse effects of the practice fall more heavily on members of the protected class than they fall on nonmembers who are similarly situated. *Steamship Clerks,* 48 F.3d at 601.

With respect to the causation prong, the Supreme Court has stated that:

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

Watson v. Fort Worth Bank & Trust, 487 U.S. 977, 994-95, 108 S.Ct. 2777, 101 L.Ed.2d 827 (1988) (plurality); *see Wessmann v. Gittens,* 160 F.3d 790, 804 (1st Cir.1998) ("Even strong statistical correlation between variables does not automatically establish causation."). The Second Circuit has summarized:

> Because statistical analysis, by its very nature, can never scientifically prove discrimination, a disparate impact plaintiff need not prove causation to a scientific degree of certainty. Accordingly, this Court has held that a plaintiff may establish a prima facie case of disparate impact discrimination by proffering statistical evidence which reveals a disparity substantial enough to
158 raise an inference of causation. That is, a plaintiff's *158 statistical evidence must reflect a disparity so great that it cannot be accounted for by chance.

EEOC v. Joint Apprenticeship Comm. of the Joint Indus. Bd. of the Elec. Indus., 186 F.3d 110, 117 (2d Cir.1999) (citations omitted). However, causation need not "invariably include a formal statistical analysis." *Steamship Clerks,* 48 F.3d at 606.

Where the employment practice at issue dispositively excludes individuals, some courts have observed that the disparate impact and causation elements appear to merge. *See Nash v. Consol. City of Jacksonville,* 905 F.2d 355, 358 (11th Cir.1990) (finding that "the fact that an examinee's failure to pass the examination absolutely bars promotion satisfies the Court's `specific causation' requirement"); *cf. Phillips v. Cohen,* 400 F.3d 388, 398 n. 8 (6th Cir.2005) ("If the employee challenges the employer's promotion process as a whole, however as is the case here then the disparate impact and causation elements merge.").

3. *Testing Caselaw*

The starting point for analysis of the Title VII claim is the seminal case *Connecticut v. Teal,* 457 U.S. 440, 102 S.Ct. 2525, 73 L.Ed.2d 130 (1982), in which the Supreme Court reaffirmed that Title VII prohibits "procedures or testing mechanisms that operate as `built-in headwinds' for minority groups." *Id.* at 448-49, 102 S.Ct. 2525 (quoting *Griggs,* 401 U.S. at 432, 91 S.Ct. 849). Commenting on congressional concern about the widespread use by state and local governmental agencies of invalid selection techniques that had a discriminatory impact, the Supreme Court stated:

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

Id. at 450, 91 S.Ct. 849. In *Teal,* the passfail examination for a public agency had a disparate impact on minorities' eligibility for promotion but no effect on the bottom-line promotion statistics because of an affirmative action program for promoting those who passed. *See id.* at 443-44, 102 S.Ct. 2525. Rejecting the agency's "bottom-line defense," the Supreme Court admonished, "[t]he suggestion that disparate impact should be measured only at the bottom line ignores the fact that Title VII guarantees these individual respondents the *opportunity* to compete equally with white workers on the basis of job-related criteria." *Id.* at 451, 102 S.Ct. 2525; *cf. Donahue v. City of Boston,* 304 F.3d 110, 119-20 (1st Cir.2002) (stating that standing to assert non-Title VII equal protection claim for prospective relief is established if plaintiff is denied "the opportunity to compete on equal footing in the[] hiring process on account of his race").

When the Supreme Court first held that the Civil Rights Act proscribes disparate impact discrimination in *Griggs,* the focus was on barriers to employment rather than who was hired.

> [T]he Act does not command that any person be hired simply because he was formerly the subject of discrimination, or because he is a member of a minority group. . . . What is required by Congress is the removal of artificial, arbitrary, and unnecessary barriers to employment when the barriers operate invidiously to discriminate on the basis of racial or other impermissible classification.

Griggs, 401 U.S. at 430-31, 91 S.Ct. 849.

159 Thus, under *Teal* and its progeny, individual components of a hiring process may *159 constitute separate and independent employment practices subject to Title VII even if the overall decision-making process does not disparately impact the ultimate employment decisions involving a protected group. *See, e.g., Stout v. Potter,* 276 F.3d 1118, 1122 (9th Cir.2002) ("The nonadverse results of the ultimate promotion decisions cannot refute a prima facie case of disparate impact at the dispositive interview selection stage."); *Smith v. Xerox Corp.,* 196 F.3d 358, 370 (2d Cir.1999) ("Even if the overall decision-making process did not create an adverse impact on a protected group, that group still has a cause of action if it can show that some component of the decision-making process caused a disparate impact."); *Newark Branch, NAACP v. City of Bayonne, N.J.,* 134 F.3d 113, 124 (3d Cir.1998) "(*Teal* suggests that a subsequent affirmative action program cannot `redeem' discriminatory conduct that produces disparate results.").

Courts have applied *Teal* to reject examinations used to rank-order candidates. For example, in *Waisome v. Port Auth. of N.Y. & N.J.,* 948 F.2d 1370, 1378 (2d Cir.1991), involving a composite score of tests to rank police officers for promotion, the Second Circuit held:

> Moreover, our prior case law lends support to the use of the [effective cutoff score]. Where a written test served, as here, both as a passing "gate" to further consideration for promotion, and as a major component of the ultimate score required for promotion, we indicated there was no disparate impact in the pass rate, but the disparity in actual promotions established that the written test had a prohibited disparate impact. In [*Kirkland v. New York State Dep't of Correctional Servs.,* 711 F.2d 1117 (2d Cir.1983)], as in the present case, evidence demonstrated that, though there was no disparity in the rate at which minority candidates for promotion passed an examination, their representation on the eligibility list was disproportionately low at the top of the list and high at its bottom. Hence, remand is required for the district court to develop a full record against which to evaluate the evidence of bunching and to determine whether the written examination had a disparate impact when these statistics and all the surrounding facts and circumstances are considered.

Id. at 1378 (citations omitted) (finding that written test served as pass-fail mechanism requiring score of sixty-six to move on in hiring process and as ranking mechanism requiring score of seventy-six to be hired, both of which constituted employment practices under *Teal*). Thus, when 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.

Teal does not necessarily require that courts deconstruct employment practices into their individual components and evaluate each for disparate impact. Recognizing the potential burden of such a requirement, some courts have interpreted *Teal* as applying Title VII protection to a component of an employment practice regardless of the "bottom-line" only if that component is an identifiable and dispositive barrier that denies an employment opportunity by preventing an individual from proceeding to the next step in the employment process. *See, e.g., District Council 37, AFL-CIO v. N.Y. City Dep't of Parks & Recreation,* 113 F.3d 347, 352-54 (2d Cir.1997) (finding that plaintiffs could challenge only "dispositive step" under *Teal* when bottom-line was nondiscriminatory); *City of*
160 *Chicago v. Lindley,* 66 F.3d 819, 829 (7th Cir.1995) (finding *Teal* applicable *160 only where "one `step' disparately excluded minority individuals from moving on to the next step and, in turn, deprived them of any *opportunity* for benefits"); *Reynolds v. Ala. Dep't. of Transp.,* 295 F.Supp.2d 1298, 1315 (M.D.Ala.2003) (finding *Teal* inapplicable because examinations did not work as absolute barrier to further hiring consideration and because there was no Title VII issue before court). Similarly, the EEOC Guidelines provide:

> C. *Evaluation of selection rates. The "bottom line". . . . If this information shows that the total selection process does not have an adverse impact, the Federal enforcement agencies,* in the exercise of their administrative and prosecutorial discretion, *in usual circukstances, will not expect a user to evaluate the individual components for adverse impact,* or to validate such individual components, and will not take enforcement action based upon adverse impact of any component of that process, including the separate parts of a multipart selection procedure or any separate procedure that is used as an alternative method of selection.

See 29 C.F.R. § 1607.4(C) (emphasis added).

4. *Statistical Evidence*

a. *Prima Facie Case*

In evaluating statistical evidence, "[t]he Supreme Court has said that no single test controls in measuring disparate impact." *Langlois v. Abington Hous. Auth.,* 207 F.3d 43, 50 (1st Cir.2000) (citing *Watson,* 487 U.S. at 995-96 n. 3, 108 S.Ct. 2777 (plurality) ("[W]e believe that such a case-by-case approach properly reflects our recognition that statistics `come in infinite variety and . . . their usefulness depends on all of the surrounding facts and circumstances'" (citation omitted).)). In this case, the plaintiffs offer the "four-fifths rule" and chi-square analysis as the statistical benchmarks.

The "four-fifths rule" comes from the EEOC's Uniform Guidelines on Employee Selection Procedures (1978) (the "EEOC Guidelines"), which provide:

> 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 by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.

29 C.F.R. § 1607.4(D). For example, if a government agency promoted female candidates 20% of the time but male candidates 40% of the time, the selection rate for females would be 50% (or half) of the selection rate for males. This 50% ratio is less than 80% and thus, would violate the four-fifths rule and demonstrate adverse impact.

The four-fifths rule is a pertinent benchmark in the employment context. *See Langlois,* 207 F.3d at 50. The Supreme Court has cautioned, however, that the rule "has not provided more than a rule of thumb for the courts." *Watson,* 487 U.S. at 995 n. 3, 108 S.Ct. 2777 (plurality) (citations omitted). Indeed, the EEOC Guidelines limit the rule's applicability in two ways. First, "[s]maller 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." 29 C.F.R. § 1607.4(D). Second, "[Oreater differences in selection rate may not constitute adverse impact where the differences are
161 based on small numbers" that "are too small to be reliable." *Id.* When *161 numbers are "too small to be reliable," the federal agencies that issued the EEOC Guidelines provided the following guidance:

> Generally, it is inappropriate to require validity evidence or to take enforcement action where the number of persons and the difference in selection rates are so small that the selection of one different person for one job would shift the result from adverse impact against one group to a situation in which that group has a higher selection rate than the other group.
>
> On the other hand, 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, warranting the need for validity evidence.

44 Fed.Reg. 11996, 11999 (Mar. 2, 1979).

At times, courts have also used chisquare calculations in determining the existence of disparate impact. *See, e.g., NAACP v. City of Mansfield, Ohio,* 866 F.2d 162, 167-69 (6th Cir.1989). Without describing the details of the calculation, chi-square is a statistical calculation that examines differences between what is expected and what is observed. For example, if a government agency hired 85 of 600 non-minority candidates (14.2% hiring rate) and 15 of 400 minority candidates (3.8% hiring rate), a chi-square calculation would indicate that the probability of the difference in hiring rates being due to chance is 0.236%. Because the probability due to chance is 0.236% and less than 5% in this example, there is "a very powerful indication that the process of selecting new hires acts disparately on [minorities] in the applicant pool." *See* Walter B. Connolly, Jr., David W. Peterson & Michael J. Connolly, *Use of Statistics In Equal Employment Opportunity Litigation* § 4.05[4][a] (1996); *see also id.* § 8.04[1][b] (applying chi-square calculation to exam scores). If the probability due to chance is 5% or less, the difference is said to rise to the 0.05 level of statistical significance (i.e., 95% certainty that difference is due to nonrandom factors). *See, e.g., Xerox,* 196 F.3d at 366-67.

b. *Job Related and Consistent With Business Necessity*

In this case, the HRD asserts that a criterion-related study validates the written cognitive examination. Criterion-related studies examine the relationship between selection procedure scores and job performance. The selection procedures are valid if "the selection procedure is predictive of or significantly correlated with important elements of job performance." 29 C.F.R. § 1607.5(B). The magnitude of the relationship is determined by calculating a correlation coefficient. *Id.* § 1607.14(B)(6).

"A correlation coefficient of + 1.0 indicates a complete identity between relative test scores and relative job performance." *Williams v. Ford Motor Co.,* 187 F.3d 533, 540 (6th Cir.1999). As Dr. Landy explained, a correlation coefficient of 0.0 indicates that the examination has no relationship with job performance. Correlation coefficients "around .30.40 are considered acceptable (by testing professionals) for tests used in selecting employees" (Ex. 24, Tab 6, at 2). *See also Beecher,* 371 F.Supp. at 516 ("[A]s a `rule of thumb' a coefficient of .3 would be the minimum level to indicate a satisfactory relationship. A lower coefficient would not be practically significant and would not justify use of the test."). Therefore, for a selection procedure to be valid under a criterion-related study, the correlation coefficient must generally exceed a 0.3 threshold.

162 In addition to the 0.3 threshold, in measuring the correlation coefficient, the calculation *162 must be "statistically significant at the 0.05 level of significance." 29 C.F.R. § 1607.14(B)(5).

Lastly, as explained by Dr. Jacobs, correlation coefficients of different selection procedures do not necessary add to one another. This is because there may be overlap when the different selection procedures correlate with overall job performance. For example, cognitive examinations and physical ability tests are both correlated with certain entry-level fire fighting duties, such as search and rescue. (*See* Ex. 1, at 14.) Therefore, the correlation coefficient of a selection procedure that uses both a cognitive examination and a physical ability test cannot be calculated by simply adding the correlation coefficients of each component.

B. *Prima Facie Case*

The plaintiffs identify the HRD's use of the entry-level civil service examination as the employment practice subject to their challenge under Title VII, asserting that it has an adverse and disparate impact with respect to the selection and consideration of minority candidates for hire as firefighters in specific municipalities and statewide. The HRD asserts that the Court should not consider the disparate impact of the examination on the scores of minorities because the plaintiffs have not demonstrated that the examination has caused a discriminatory impact on the hiring of minorities.

1. *Examination Scores*

The statistical evidence shows clear differences in scores as a function of race for both the 2002 and 2004 examinations.[7] On average, minority candidates score lower than non-minority candidates:

	Whites	African Americans	Hispanics
2002 Examination	89.10	77.80	78.44
2004 Examination	88.10	77.00	78.70

(Ex. 33A.) The examination scores of minority candidates are disproportionately lower at all scores above the nominal passing score of seventy. (*See* Ex. 1, Tables 6-7.) In addition, all parties agree that to be hired in most communities, which are not applying for a special certification (e.g., Spanish-language ability) or subject to the *Beecher* certification quotas, non-veteran candidates must obtain a score above an effective cutoff score of ninety. (Ex. 24, at 68-69.) In other words, in most municipalities, any non-veteran candidate with a score under ninety who is hired is a rare bird. Accordingly, the following table demonstrates the passing rates and ratio of passing rates for the 2002 examination at the scores of seventy, which is the nominal passing score, and ninety, which is the effective cutoff score for non-veteran candidates in most communities.

	Ratio of African American	Ratio of Other Minority

2002 Score	Majority Passing Rate	African American Passing Rate	and Majority Passing Rates	Other Minority Passing Rate	and Majority Passing Rates
90	60.20%	19.93%	33.11%	22.63%	37.60%
70	96.20%	76.20%	79.19%	77.96%	81.03%

163 *163 (*See* Ex. 1, Table 6; Ex. 10, Attach. I.)[8] For the 2004 examination:

2004 Score	Non Protected Group Passing Rate	African American Passing Rate	Ratio of African American and Non-Protected Group Passing Rates	Hispanic Passing Rate	Ratio of Hispanic and Non-Protected Group Passing Rates
90	52.47%	17.69%	33.71%	19.12%	36.45%
70	96.52%	75.55%	78.27%	80.36%	83.26%

(*See* Ex. 1, Table 7; Ex. 12, Attach. G.)

Under the four-fifths rule, there is no adverse impact on minorities (Hispanics and African-Americans) at the nominal passing score of seventy. (*See* Trial Tr. 39-40, Apr. 14, 2006.) It is undisputed, however, that the four-fifths rule is violated for every score greater than seventy. (*See* Ex. 1, Tables 6-7.) The examinations thus have a severe adverse and disparate impact on non-veteran minority candidates, who must score above the effective cutoff score of ninety to be hired in most communities. The chi-square calculations support these findings by demonstrating that the differences in scores by race are statistically significant. (*See* Ex. 1, Tables 8-9.)

2. *Hiring Statistics*

Examination statistics are not determinative of the critical issue of whether the examination disparately precludes minority candidates from being hired. The plaintiffs argue that the 2002 and 2004 examinations have an adverse and disparate impact on the employment opportunities of these minority candidates, as demonstrated by an overall comparison of the percentage of candidates who passed the examination to the percentage of those hired. In rebuttal, the HRD urges the Court to take a more nuanced approach, insisting that these overall statistics improperly aggregate numbers statewide and gloss over the major impact of preferences for veterans and residents on minority hiring. The HRD believes that adverse impact must be determined by municipality because each is a separate hiring unit with different requirements and separate appointing authorities.

The Court finds merit in some of the HRD's arguments. The statutory framework in Massachusetts gives candidates with veteran status priority over those without, regardless of examination score. This distinction between veterans and nonveterans is important because there is a lower percentage of minorities in the veteran
164 candidate pool. *164

	2002 Exam	2004 Exam

Minority Candidates With Veteran Status	83	78
Minority Candidates Total	1845	864
% of Minority Candidates With Veteran Status	4.5%	9.0%
Non-Minority Candidates With Veteran Status	824	822
Non-Minority Candidates Total	10052	6165
% of Non-Minority Candidates With Veteran Status	8.2%	13.3%

(Ex. 331, at 5-6.) Therefore, to ensure that apples are compared with apples, the Court evaluates veterans and non-veterans separately.[9]

In addition, the Court finds that the statewide aggregated approach advocated by the plaintiffs and the disaggregated municipality approach advocated by the HRD are both useful in this case. As Dr. Jacobs observes, determining the appropriate level of hiring analysis "is a complicated question that statisticians would argue about, but my opinion is that when we look at what happens in hiring, we try and replicate the decision process." (Trial Tr. 100-02, Apr. 13, 2006.) The decision process in hiring firefighters in Massachusetts, however, contains multiple steps that span both the state and municipal levels. *See* *Bradley v. City of Lynn,* 403 F.Supp.2d 161 (D.Mass.2005) (holding HRD to be Title VII employer). Therefore, the Court considers hiring statistics on both statewide and municipality levels.

Lastly, the Court does not compare hiring data from municipalities subject to the *Beecher* certification quotas with those that have been released from the decree. Indeed, the parties focused on the hiring statistics of only non-*Beecher* municipalities, and I do so as well.

a. *Statewide Aggregated Hiring Data*

i. *Candidates With Veteran Status*

The following table summarizes the statewide veteran hiring statistics for non-*Beecher* municipalities, excluding data from municipalities that did not appoint anyone or had no minority candidates.

	Number of Candidates Taking Examination	Number Appointed	Selection Ratio
Minorities (2002)	37	4	10.8%
Non-Minorities (2002)	187	67	35.8%
Minorities (2004)	45	12	26.7%
Non-Minorities (2004)	275	127	46.2%

(*See* Ex. 33E, Table 2; Ex. 33F, Table 2.) Under the four-fifths rule, the disparate impact ratio is 30% for the 2002 examination and 58% for the 2004 examination, evidencing adverse impact (i.e., less than 80%). The chi-square calculations support these findings by demonstrating that the differences in scores by race are statistically significant. (*See* Ex. 33E, at 2; Ex. 33F, at 3.)

ii. *Candidates Without Veteran Status*

The following table summarizes the statewide non-veteran hiring statistics for non-*Beecher* municipalities,
165 excluding data *165 from municipalities that did not appoint any non-veterans.

	Number of Candidates Taking Examination	Number Appointed	Selection Ratio
Minorities (2002)	210	6	2.9%
Non-Minorities (2002)	2561	152	5.9%
Minorities (2004)	62	3	4.8%
Non-Minorities (2004)	456	41	9.0%

Under the four-fifths rule, the disparate impact ratio is 49% for the 2002 examination and 54% for the 2004 examination, evidencing adverse impact (i.e., less than 80%). (*See* Ex. 5, Table 3; Ex. 6, Table 3.)

b. *Disaggregated Hiring Statistics*

The HRD argues that statistical evidence aggregated on a statewide basis is flawed because hiring decisions are made individually by each municipality at the local level and because a large municipality can distort the whole adverse impact analysis. Dr. Jacobs asserts that aggregation "allows one single jurisdiction to have a major influence on the entire system leading to a conclusion of adverse impact for the entire Commonwealth when a single community is responsible for the outcome." (Ex. 33S.) Specifically, Dr. Jacobs argues that Boston's hiring distorts the calculation, because if it is remedied, the state passes the four-fifths rule.

Because this seems reasonable, the Court examines the disaggregated hiring data of Boston. The Court also examines the disaggregated hiring data of Lynn, the municipality of the four named plaintiffs. The data shows that the examination has a disparate impact on the hiring of minorities in both of these municipalities.

i. *Boston*

The following table summarizes the Boston hiring statistics for the 2004 examination of candidates with veteran status.[10]

	Non-Minorities	Minorities
Number of Candidates	177	27
Number Appointed	90	7
Appointment Ratio	51%	26%

(*See* Ex. 33D, at 2.) Under the four-fifths rule, the disparate impact ratio is 51%, evidencing adverse impact (i.e., less than 80%). The chi-square calculations support these findings by demonstrating that the differences in scores by race are statistically significant. (*See id.*)

It is also important to note that Boston has hired two separate classes of entry-level firefighters from the 2004 examination one in June 2005 and one in January 2006. The following table summarizes in which class candidates were hired as a function of race.

	June 2005 Class	January 2006 Class
Non-Minorities	50	40
Non-Minorities Hired in Class/ Non-Minorities Hired Total	55%	45%
Minorities	2	5
Minorities Hired in Class/ Minorities Hired Total	28%	72%

166 *166 (*See* Ex. 33D, at 3.) Under the four-fifths rule, the disparate impact ratio for the first June 2005 class is 51%, evidencing adverse impact (i.e., less than 80%). Stated plainly, white veterans were twice as likely to be appointed in the first June 2005 class as minority veterans. (*Id.*)

ii. *Lynn*

The data from Lynn, where the named plaintiffs reside, are less clear. Since 1986, when Lynn was released from the *Beecher* decree, it is estimated that only four of the 106 entry-level firefighters have been minorities. The following table summarizes the Lynn non-veteran hiring statistics for the 2004 examination.

	Non-Minorities	Minorities
Number of Candidates	69	15
Number Appointed	12	0
Appointment Ratio	17.4%	0.0%

(*See* Ex. 33, at 17.)[11]

As a statistical matter, the experts disagree on whether a disparate impact ratio can be calculated under the four-fifths rule when Lynn hired zero minorities. That may be an interesting mathematical quandary, but in this case, the "fine tuning of the statistics could not have obscured the glaring absence of minority" hires. *Int'l Bhd. of Teamsters v. United States,* 431 U.S. 324, 342 n. 23, 97 S.Ct. 1843, 52 L.Ed.2d 396 (1977). The Court concludes that the 2004 examination has had a disparate impact on the hiring of non-veterans in Lynn.

However, the record is poor on the impact of the 2004 examination on veterans. The record does not contain a summary of Lynn hiring for the 2004 examination of candidates with veteran status. Thus, the record is too sparse to tell if there has been a disparate impact on hiring veteran minority candidates in Lynn from the 2004 examination.

The record is equally unclear with respect to all candidates from the 2002 examination. According to Dr. Jacobs' summary statistics, Lynn hired eight candidates from the 2002 examination; all were paramedics without veteran status, but none were minorities. (Ex. 32, at 15.) Since all were non-veteran paramedics, this would appear to be a class hired pursuant to a special certification request for emergency medical technicians. However, nothing in the record indicates whether the class was the result of a special certification, and the raw certification data is again unhelpful as it conflicts with Dr. Jacobs' summary data. According to the raw certification data, Lynn hired twelve candidates from the 2002 examination, some of whom were not veterans or paramedics. (*See* Ex. 17.) Thus, the record is too sparse to evaluate the disparate impact on hiring in Lynn from the 2002 examination.

iii. *Other Municipalities*

The HRD contends that in most communities, the number hired is too small to determine whether there has been a disparate impact under the four-fifths rule or to be statistically significant. Disparate impact in hiring is not found in some municipalities because the numbers are "too small to be reliable" under the four-fifths rule. (*See* Ex. 331, at 3-4.) For example, if Haverhill had hired one additional minority from the 2002 examination, its hiring statistics would change from the presence to the absence of adverse impact under the four-fifths rule. (*See id.* at 3
167 (showing minority shortfall under four-fifths *167 rule to be less than one for Haverhill).)

While it is true that in some municipalities there is no disparate impact, based on the testimony of Dr. Landy and Dr. Wiesen, the plaintiffs have demonstrated a reasonable statistical basis to address the "small numbers" problem by aggregating non-*Beecher* communities with minority applicants to determine whether there was an adverse impact on minority hiring. Given the statutory framework mandating the HRD's involvement in the hiring process across municipalities, the aggregation approach is supported in this case by the EEOC Guidelines and the caselaw. 29 C.F.R. § 1607.4(D);[12] *Vulcan Pioneers, Inc. v. N.J. Dep't of Civil Serv.,* 625 F.Supp. 527, 534-35, 544-45 (D.N.J.1985) (finding aggregation across municipalities and across years appropriate where State administered firefighter promotion examination for municipalities and examinations were extremely similar across years). *But see* Bailey v. Se. Area Joint Apprenticeship Comm., 561 F.Supp. 895, 901, 910 (N.D.W.Va.1983) (finding aggregation across localities inappropriate where each jurisdiction implemented standards autonomously); *but cf.* Fudge v. City of Providence Fire Dep't, 766 F.2d 650, 656-57 (1st Cir.1985) (finding aggregation across years clearly erroneous where examinations were sufficiently different). Moreover, even examining Boston and Lynn independently at the municipality level, there is a disparate impact from the 2004 examination in hiring veterans in Boston and nonveterans in Lynn.

3. *Causation*

The plaintiffs have thus demonstrated through significant statistical evidence not only that the examination has a disparate impact on the scores of minority candidates but also that there is a disparate impact on the hiring of minorities regardless of veteran status statewide, of minority veterans in Boston, and of minority non-veterans in Lynn. Coupled with the statistical evidence is the fact that the statutory framework by ranking candidates by score makes the examination integral to whether and when individuals are hired. For example, Jacob Bradley, one of the named African American plaintiffs, was neither hired nor certified for hiring consideration in Lynn in 2005 because of his examination score. Despite scoring *ninety-four* on the 2004 examination, Jacob Bradley scored too low. In contrast, a score of *ninety-five* enabled three white non-veteran candidates to be certified for hiring consideration, and Lynn hired all three. (*See* Exs. 14, 18; Trial Tr. 20, Apr. 12, 2006.)

Significantly, Dr. Jacobs, the HRD's expert, conceded that "race does have some adverse impact in the hiring process in Massachusetts." (Ex. 26, at 4; Trial Tr. 121-22, Apr. 14, 2006.) Dr. Jacobs added that, based on his experience as an expert in the field, jurisdictions that use cognitive ability examinations as the sole basis for rank ordering are "likely to get adverse impact." (Trial Tr. 100, Apr. 14, 2006.)

The HRD asserts several arguments in rebuttal regarding causation.

168 *168 a. *Statutory Preferences*

First, the HRD argues that the statutory preference for veterans causes the disparate impact in hiring. Dr. Jacobs testified: "Well, the chain of evidence that I see is that minorities score lower on the test. Minorities have less

veterans status. And those two things combine to have whatever the adverse impact ratios might be, whatever way we calculate it. What proportion is in the test, what proportion is to the veterans status I can't tease out." (Trial Tr. 122, Apr. 14, 2006.) While it is true that the veterans preference has a disparate impact on minority hiring, the use of the examination also has a disparate impact on hiring even within the veterans category.

To be sure, this disparate impact does not exist in every municipality. For example, municipalities such as Haverhill had only one vacancy, and the veteran candidates on the certification list had scores ranging from 81 to 95. All the candidates on the list were white, and there is no evidence there were any minority veteran candidates who resided in Haverhill. (*See* Ex. 21, at 11.) Therefore, in Haverhill, the lower minority examination scores did not cause the hiring disparity.

To address the HRD's concern that the examination played no role in the hiring in some municipalities, the Court certified a class pursuant to Fed.R.Civ.P. 23(b)(2) for liability only. With respect to remedies for individual applicants, the Court will have to look separately at each municipality to determine whether there were any minority applicants in the hiring category impacted by the examination. The differences in the demographics of each municipality, however, do not undermine a finding of a prima facie case, which need not be proven to a mathematical certainty.

b. *Timing*

Next, the HRD argues that it is too early to assess whether there was adverse impact in hiring from the 2004 examination because some municipalities are still appointing from the 2004 list. In a municipality like Boston, which hires in multiple classes, the delay in hiring alone constitutes adverse and disparate impact. Municipalities "may appoint only from among the first 2n + 1 persons named in the certification willing to accept appointment." (Ex. 9, at 16-17 (Pers. Admin. Rule 9(1)).). Thus, regardless of whether an applicant is a veteran or non-veteran, the examination score determines both whether and *when* a candidate is certified or hired. For example, veteran candidates had no opportunity to be hired in Boston's June 2005 class unless they were ranked higher on the certification list than a score of ninety-three (*see* Ex. 21, at 2-7). *Cf.* *Waisome,* 948 F.2d at 1377-78 (identifying "the minimum score a candidate could achieve on the written component and still be within" the candidates actually hired to be determinative of whether a candidate had a "real opportunity for promotion"); *Guinyard v. City of New York,* 800 F.Supp. 1083, 1088-89 (E.D.N.Y.1992) (finding that delay in promotions of minority candidates to police captain may constitute adverse and disparate impact).

The effect of using examination scores, which disparately impact minorities at all scores above seventy, for rank ordering, is to bunch minorities at the bottom of the eligible list. Of the last seven veterans on the 2004 Boston list, five are minorities. (Ex. 33H.) Ranking by examination score thus disproportionately has precluded minority candidates from hiring consideration in the initial June 2005 class. Even if hired in future classes, minorities as a
169 class have been adversely and disparately *169 impacted by loss of pay, benefits, and seniority caused by the delay.

c. *"Drop Out" Rate*

Lastly, the HRD argues that the "drop out" rate of minority applicants exceeds that of majority applicants, and that this causes the disparate hiring statistics. The plaintiffs dispute that the disparity exists or is significant.

To determine the "drop out" rate, Dr. Jacobs identified the lowest score hired on the ranked certification list. All candidates on the list with a score at or above the lowest score hired were candidates reached for consideration. Those reached for consideration but not hired were candidates who "dropped out." The "drop out" rate is the percentage of candidates who "dropped out" among those who were reached for hiring consideration. (*See* Ex. 33S, at 7.) Therefore, based on the statutory and administrative framework, the "drop out" rate used by the HRD included not only candidates who voluntarily were unwilling to serve but also those whom municipalities bypassed for various reasons, including a drug test, background check, physical abilities test, or medical or psychological examination.

On a statewide aggregate level, Dr. Landy found that the "drop out" rate for minorities and non-minorities was essentially the same—48% for minorities and 47% for whites. (Ex. 33T, at 2.) Dr. Jacobs did not disagree. (*See* Ex. 33U, at 3.)

The bone of contention was Boston. Dr. Landy and Dr. Jacobs engaged in an ongoing point-counterpoint dialogue during and after trial to determine whether there was a difference in "drop-out" rates between minorities and non-minorities in Boston. In the end, Dr. Landy and Dr. Jacobs agreed that in Boston the "drop out" rate for minorities was higher than the rate for whites, but disagreed on the scope and significance of the difference. Dr. Jacobs believed that the "drop-out" rate was 58% for minorities and 41% for whites in Boston and that the difference was a "big difference." Dr. Landy believed that the "drop-out" rate was 53% for minorities and 41% for whites and that the difference was not statistically significant.

Unfortunately, the HRD did not address "drop-out" rates until late in the trial, and neither expert focused on these rates in their original expert reports or was subject to cross-examination.[13] While the statistical evidence was heavily disputed, I conclude that the higher minority "drop out" rate in Boston does not undercut the plaintiff's prima facie case.

The only difference in "drop out" rates occurred in Boston, which hired in at least two stages from the 2004 examination— June 2005 and January 2006. Dr. Jacobs did not differentiate between the "drop out" rates in these different classes. Neither did he evaluate the particular factors causing the "drop out." There is nothing in the record to demonstrate when a minority candidate "dropped out" or the reason for the "drop out." Boston had a pool of approximately 212 veteran candidates from the 2004 examination. For the first June 2005 class, however,
170 only the top 109 were reached for consideration. (*See* Ex. *170 21, at 2-7.) Thus, the "drop out" rate had no effect on the disparate impact on minority veterans who scored too low even to be reached for hiring consideration in June 2005.

Moreover, rank ordering by written cognitive examination likely caused the "drop out" rate to be higher for minority candidates. Minority veteran candidates were bunched by cognitive examination score at the bottom of the veterans list and were not reached for hiring consideration until the second January 2006 class, or later. It is reasonable to infer that more minority candidates were unwilling to serve and thus, "dropped out" because they took other employment opportunities with the passage of time and the need to earn a living. Based on the literature in the field, Dr. Landy points out:

> [T]he length of time an applicant has to wait before appointment (white or minority) has a substantial impact on drop out likelihood—the longer you wait for appointment, the more likely it is that you will have taken another job or lost interest in the job in question if and when the employment offer comes.

(Ex. 33T, at 3.) Therefore, because the difference in examination scores causes a difference in the timing of employment opportunities, the higher minority "drop out" rate in Boston does not undermine the finding of disparate impact and causation in Boston.

As stated by the First Circuit in *Beecher,* the purpose of a prima facie determination is to ferret out qualifications that are reasonable to ask an employer to justify. "When widespread minority underemployment is shown to exist in a given occupation, primary selection devices should not be immunized from study by placing unrealistically high threshold burden upon those with least access to relevant data." *Beecher,* 504 F.2d at 1020-21. Accordingly, the Court concludes that the statistical evidence is substantial, significant, and sufficient to raise an inference of causation, and the plaintiffs have demonstrated a prima facie case here.

C. *Job Related and Consistent With Business Necessity*

Once the plaintiffs prove a prima facie case that withstands the head-on attacks by the defendants, the burden of production and persuasion shifts to the defendants to prove that "the challenged practice is job related for the position in question and consistent with business necessity." 42 U.S.C. § 2000e-2(k)(1)(A)(i), (m); *see Steamship Clerks,* 48 F.3d at 601-02.

One of the explicit purposes of the Civil Rights Act of 1991 is "to codify the concepts of `business necessity' and `job related' enunciated by the Supreme Court in *Griggs v. Duke Power Co.,* 401 U.S. 424, 91 S.Ct. 849, 28 L.Ed.2d 158 (1971), and in the other Supreme Court decisions prior to *Wards Cove Packing Co. v. Atonio,* 490 U.S. 642, 109 S.Ct. 2115, 104 L.Ed.2d 733 (1989)." Pub.L. No. 102-166 § 3(2). In *Griggs,* the Supreme Court

found a Title VII violation because the employment practices disparately impacted the plaintiffs based on race and because "neither the high school completion requirement nor the general intelligence test [wa]s shown to bear a demonstrable relationship to successful performance of the jobs for which it was used." *Id.* "Congress has placed on the employer the burden of showing that any given requirement must have a manifest relationship to the employment in question." *Griggs,* 401 U.S. at 432, 91 S.Ct. 849.

171 To pass muster under Title VII, the civil service examination must be both "job related" for the entry level fire *171 fighter position and consistent with "business necessity." *Langlois,* 207 F.3d at 53-54 (stating that employment practice may be discriminatory without intent if it, "without demonstrably advancing the interest asserted in justification, somehow impedes persons of color from competing on an equal footing with others"). To be sure, "employers are not required, even when defending standardized or objective tests, to introduce formal `validation studies' showing that particular criteria predict actual on-the-job performance." *Watson,* 487 U.S. at 998, 108 S.Ct. 2777 (plurality).

> In *Beazer,* for example, the Court considered it obvious that "legitimate employment goals of safety and efficiency" permitted the exclusion of methadone users from employment with the New York City Transit Authority ..... 440 U.S. at 587, n. 31[, 99 S.Ct. 1355]. Similarly, in *Washington v. Davis,* the Court held that the "job relatedness" requirement was satisfied when the employer demonstrated that a written test was related to success at a police training academy "wholly aside from [the test's] possible relationship to actual performance as a police officer." 426 U.S. at 250[, 96 S.Ct. 2040].

Watson, 487 U.S. at 998-99, 108 S.Ct. 2777 (plurality). The HRD, however, must do more than ask the Court "to undertake a leap of faith. . . . If courts were to accept an employer's arbitrary *ipse dixit* as a satisfactory justification for retaining a policy that produces an invidiously discriminatory impact, Title VII would be reduced to no more than a toothless tiger." *Steamship Clerks,* 48 F.3d at 607.

The HRD asserts that it has shouldered its burden under Title VII through the 1992 Report (Ex. 27), which was conducted in accordance with the EEOC Guidelines. As such, the Court looks to the EEOC Guidelines, as "a body of experience and informed judgment to which courts and litigants may properly resort for guidance." *Mentor Say. Bank v. Vinson,* 477 U.S. 57, 65, 106 S.Ct. 2399, 91 L.Ed.2d 49 (1986). The EEOC Guidelines describe three types of validity studies criterion-related, content, and construct all of which first require a job analysis. *See* 29 C.F.R. §§ 1607.5, 1607.14(A). Based on the job analysis, the EEOC Guidelines detail the minimum technical standards regulating topics such as the representativeness of the sample, statistical significance, and reliability for validating the selection procedure in the three types of studies. *See id.* § 1607.14.

In addition, the "evidence of both the validity and utility of a selection procedure should support the method the user chooses for operational use of the procedure." *Id.* § 1607.5(G). To validate the use of examinations for ranking:

> 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. Thus, if a user decides to use a selection procedure on a ranking basis, and that method of use has a greater adverse impact than use on an appropriate pass/fail basis . . ., the user should have sufficient evidence of validity and utility to support the use on a ranking basis.

Id. In this case, the "four-fifths rule" statistics demonstrate that the use of the examination for ranking has a greater adverse and disparate impact than the use of the examination for pass/fail. Therefore, it is not enough to validate the examinations generally. To pass muster under Title VII and the EEOC Guidelines, the HRD must
172 validate the scoring on examinations to support their use on a ranking *172 basis. *Cf. Bew v. City of Chicago,* 252 F.3d 891, 894-95 (7th Cir.2001) (finding that examinations "must be scored so that it properly discriminates between those who can and cannot perform the job well" (citation and quotations omitted)).

The 1992 Report documents the validity of the written cognitive examination in Massachusetts by showing that criterion-related evidence from a 1986 Columbus, Ohio study could be appropriately transported here. The 1995

Report and the 2002 Job Analysis both document professionally-conducted analyses of the tasks and abilities of entry-level firefighters; however, all three experts agreed that neither of these analyses documented a validity study.[14] The plaintiffs thus assert that the validity evidence is outdated because the only proffered validity study comes from the 1992 Report. The plaintiffs also assert that the 1992 Report does not support the use of the written cognitive examination as the sole basis for ranking. Nonetheless, the HRD asserts that the 1992 Report is sufficient because the 1995 Report and 2002 Job Analysis show no evidence of the job duties of entry-level firefighters changing between 1992 and 2004. The Court finds that the HRD has failed to meet its validation burden for several reasons.

First, the HRD's own expert recommends, and other experts agree, that a validity study should be conducted every five years. The 1992 Report is not only a decade old but relies on 1986 data from another jurisdiction. This is too long a hiatus under the standards in the industry. Second, the 1992 Report validates an examination professionally-written by experts in industrial psychology and test development. By contrast, the 2002 and 2004 examinations, which were written by the HRD based on past examinations, are neither professionally-created nor professionally-validated. *See* *Beecher,* 504 F.2d at 1024-25 (affirming that validation study of examination did not survive Title VII scrutiny in part because examination was not professionally developed). The HRD explains that professionals were unnecessary because past examinations served as models and that there is no evidence that the 2002 and 2004 examinations differed significantly from past examinations. However, confidence in the validity of the examination is undermined by the evidence that the HRD adjusted scores after administering the examination by removing questions and by crediting multiple answers as correct on questions so that the arbitrary passing point of seventy produced no adverse impact on minorities under the four-fifths rule.[15] Based on the reliance on an outdated validity study, the failure to use professional experts, and the practice of gerry-rigging the test, I find that the HRD dropped the *Beecher* ball and failed to ensure that the 2002 and 2004 examinations were properly validated.

Even more significantly, the 1992 Report validated rank ordering by score only when a cognitive test constituted 40% and a physical test constituted 60% of the overall composite score. Dr. Landy, who supervised the 1992 Report, testified to this, and this composite score was what the HRD implemented for the hiring process that
173 followed the 1992 Report. Indeed, in *173 evaluating the same 1986 criterion-related study from Columbus, Ohio, which the 1992 Report was based upon, the Sixth Circuit stated: "We reiterate that a selection procedure that ranks only on the basis of [cognitive ability test] scores is not acceptable." *Brunet v. City of Columbus, Ohio,* 58 F.3d 251, 255 (6th Cir.1995). The HRD has never validated the use of the written cognitive examination as the sole basis for rank ordering and thus, has failed to meet its burden. 29 C.F.R. § 1607.5(g) ("[T]he user should have sufficient evidence of validity and utility [in all studies under the EEOC Guidelines] to support the use on a ranking basis."); *id.* § 1607.14(B)(6) ("Users should evaluate each selection procedure [in a criterion-related study] to assure that it is appropriate for operational use, including establishment of cut-off scores or rank ordering.").

While cognitive ability examinations predict overall entry-level firefighter job performance to some degree (*see* Ex. 1, at 14), *both* Dr. Jacobs and Dr. Landy testified that the 2002 and 2004 examinations cannot be used reliably to distinguish candidates within a spread of as much as eight points. (Trial Tr. 15-16, Apr. 12, 2006 (Dr. Landy: "[T]he margin of error is 8 points. That would be considered confidence interval. So that there would be no difference between a score of 100 and a score of 92."); Trial Tr. 8, Apr. 13, 2006 (Dr. Jacobs: "[Dr. Landy]'s about right. It might be a little smaller because the test reliability was .9, which means that spread might be six points. But there is a spread ..... Where it doesn't matter.").)

In light of the evidence that cognitive abilities have a relatively low correlation with overall job performance (a correlation coefficient of between 0.2 and 0.3) and this eight-point margin of error, nothing in the record supports the HRD's stand-alone use of the written cognitive examinations to distinguish and rank candidates by single examination points. *Cf.* *Boston Police Superior Officers Fed'n v. City of Boston,* 147 F.3d 13, 23-24 (1st Cir.1998) (finding that one-point difference in promotional exam was "as a matter of testing accuracy, negligible" where evidence showed that "candidates who scored within a three-point band should be considered functionally equivalent . . . and equally qualified to successfully perform the job as any other person in that score band"); *Kirkland v. N.Y. State Dep't of Correctional Serv.,* 711 F.2d 1117, 1133 (2d Cir.1983) (accepting zone-banding

approach to "eliminate[] the central cause of the adverse impact" and to "create[] a more valid method to assess the significance of test scores").

To summarize, as all experts testified, I find that cognitive ability is correlated with job performance in public safety positions and thus, that cognitive ability examinations, in part, predict entry-level firefighter job performance. However, these cognitive examinations do not predict how quickly a firefighter can climb stairs with equipment or raise a ladder. Memorization skills only carry you so far. Teamwork and physical prowess are even more highly correlated with job performance. There is no persuasive evidence in this record that the use of the written cognitive examination as the *sole* basis for rank ordering entry-level firefighter candidates is a valid selection procedure.

Thirty years after *Beecher,* the First Circuit's words regarding the validity of the civil service examination unfortunately continue to ring true: "Too many doubts persist concerning the validity of this test, the format of which has persisted for years, to make a convincing case for its unaltered use in fire departments notable for the absence of minority employees." 504 F.2d at 1022. Accordingly, the Court concludes that the HRD has not met its
174 *174 burden of demonstrating that use of the 2002 and 2004 civil service examinations for rank ordering is job related and consistent with business necessity.

D. *Alternative Employment Practices*

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. While there is no dispute that cognitive ability examinations provide information relevant to the selection of entry-level fire-fighters, the experts from both sides discussed several acceptable alternatives.

First, the HRD could have banded the written cognitive examination scores. As all experts agree, there is no rational, statistically valid basis for distinguishing between candidates within a band of eight points because of the examination's reliability. A score of 100 is thus no different from a score of 92 in predicting job performance. The HRD has expressed some legal uncertainty as to whether the statutory framework in Massachusetts allows banding. The statute does require that candidates in each category "shall be arranged . . . in the order of their marks on the examination." Mass. Gen. Laws ch. 31, §§ 25-26. The Personnel Administration Rules state that the HRD shall "certify the names standing highest on such list in order of their place." (Ex. 9, at 14.) While the attorneys have not briefed the issue, banding based on scores that have no statistical difference to diminish the adverse impact of a rank-order system seems consistent with the statutory scheme and applicable caselaw under Title VII. Other courts have adopted the banding approach to "eliminate[] the central cause of the adverse impact, i.e., the rankordering system" and to "create[] a more valid method to assess the significance of test scores." *Kirkland,* 711 F.2d at 1133.

Second, the HRD could have used a physical abilities, personality (a.k.a. work style),[16] and/or biodata (a.k.a. life experience)[17] test in combination with the written cognitive examination to rank candidates. Based on the record, the HRD is one of the few major jurisdictions nationwide that uses a written cognitive examination as the exclusive basis for ranking firefighter candidates. Other major jurisdictions use written cognitive examinations either for pass/fail purposes only or for ranking in conjunction with these other tests for an overall score.

While none of these approaches alone provides the silver bullet, these other noncognitive tests operate to reduce the disparate impact of the written cognitive examination. Dr. Landy states that the "scientific literature clearly illustrates the efficacy of such an approach" (Ex. 1, at 71-74), and Dr. Jacobs agrees that such an approach "would reduce the adverse impact from what it is now to a much better level." (Trial Tr. 124-25, April 14, 2006.)

175 *175 In addition, the use of non-cognitive tests with the written cognitive examination increases the validity of the selection procedure. Statistically speaking, incorporating physical, personality, and/or biodata into the ranking mechanism increases its correlation coefficient. For example, the written examination that Dr. Jacobs currently recommends to his clients tests for cognitive, personality, and biodata. The written cognitive examination alone has a correlation coefficient of 0.2 to 0.3. By adding the personality and biodata tests to the examination, the correlation coefficient of Dr. Jacobs' examination increases to a range of 0.3 to 0.35. In Dr. Jacobs' words, the

correlation coefficient goes "up substantiqlly and I can use this term in this case `significantly.'" (Trial Tr. 86-87, Apr. 14, 2006.) Similarly, a physical test alone has a correlation coefficient between 0.3 and 0.4, making it a better predictor than cognitive examinations. The experts agree that there is a higher correlation between job duties and physical ability than between job duties and cognitive ability. *See also* *Zamlen v. City of Cleveland,* 906 F.2d 209, 217-20 (6th Cir.1990) (holding that rank ordering by combination of cognitive and physical test is valid and job related under Title VII even though physical test may have adverse impact on women). Plainly speaking, combining the physical, personality, and/or biodata tests with the written cognitive examination would not only allow the HRD to continue to serve its interest in selecting candidates based in part on cognitive ability but would also make the rank ordering of candidates by score a better predictor of overall entry-level firefighter performance.

The HRD argues that this multipronged testing approach to establish a candidate's ranking fails because the plaintiffs have not proposed precise personality or biodata questions, grading mechanisms for a physical abilities test, or weights of the different tests. The plaintiffs, however, have no obligation to provide the exact floor plan. The 1992 Report sets forth a specific alternative: ranking based on 60% physical and 40% cognitive. The Court finds that the HRD too quickly deep-sixed the use of the physical test as a significant component of the ranking, as originally recommended by Landy-Jacobs. Dr. Jacobs testified that most fire departments consider physical ability to comprise about fifty percent of a firefighter's job, yet in Massachusetts, it has no role in ranking candidates. While the HRD did have medical problems with the initial physical test, those problems were solved.

Accordingly, the plaintiffs have demonstrated the availability of several alternative selection procedures. While there is no cookie-cutter approach to hiring firefighters in the field, other jurisdictions have managed to devise selection procedures to have a less adverse impact. Massachusetts has had over thirty years to fine-tune a better approach, but the *Beecher* certification quotas provided a convenient shortcut inducing the HRD to forget the other *Beecher* mandate to create a better examination.

E. *Beecher Decree*

The plaintiffs also assert that the HRD has violated the *Beecher* decree. The First Circuit affirmed the district court's decision that the entry-level firefighter examination given at that time had not been shown "demonstrably [to] select people who will perform better the required on-the-job behaviors after they have been hired and trained." *Beecher,* 504 F.2d at 1021-1022. Specifically, the First Circuit stated:

> Too many doubts persist concerning the validity of this test, the format of which has persisted for
176 years, to make a convincing *176 case for its unaltered use in fire departments notable for the absence of minority employees. Although perfect tests are goals as illusory as perfect schools . . . the evidence justifies compelling defendants to attempt to fashion a more sensitive test, one that will not needlessly serve as a "built-in head" wind to competent minority members, depriving both them and the Commonwealth of an opportunity for which they are qualified.

Id. at 1022. Thirty years later, not much has changed.

The, *Beecher* decree orders the HRD to "cease using written firefighter entrance examinations" and provides that "[s]hould the [HRD] desire to utilize entrance examinations in the future for the purpose of selecting firefighters, such examinations shall be demonstrably job-related and validated in accordance with the [EEOC Guidelines], or otherwise shown to have no discriminatory impact." 371 F.Supp. at 521. As discussed under the Title VII analysis, the HRD has failed to show that the 2002 and 2004 examinations are validated, and the plaintiffs proved discriminatory impact. Accordingly, the Court finds that the HRD has violated its obligations under the *Beecher* decree.

The HRD first responds that the plaintiffs have no standing as non-parties to enforce the decree. The class plaintiffs acknowledge that "they are not parties to the original decrees." However, Intervenor Plaintiff NAACP is one of the parties who brought the original *Beecher* lawsuit against the HRD and is a party to the decree. *See* 371 F.Supp. at 510.

Next, the HRD argues that the Beecher decree was intended to remedy the effects of only past discrimination and thus, confers benefits only to minorities in communities where parity has not been attained. However, the

provision requiring the HRD to validate its examinations is a separate, independent, and continuing obligation. The district court "retain[ed] jurisdiction for such further action as may be necessary or appropriate" and stated that the "decree is subject to amendment where the parties so agree and with the approval of the Court." *Beecher,* 371 F.Supp. at 520, 523. To be sure, "the decree was not meant to operate in perpetuity." *Quinn,* 325 F.3d at 24. Therefore, once the examination is finally fixed, and the effects of any discrimination cured, it might be time to dissolve the decree. Until that time, it must be followed.

The defendants argue, however, that timing also cuts against the plaintiffs. The *Beecher* decree states:

> If the parties disagree as to whether a written examination has been shown to be valid within the meaning of the Guidelines, the question of their validity and job relatedness shall be resolved by the Court, and such resolution, whether by the parties' agreement or by the Court, shall be accomplished *before any such test is put into use* for the purpose of qualifying or selecting.

Id. (emphasis added). The plaintiffs have questioned the consistency of the HRD's Goldilocks position that it is "too early" to challenge the 2004 examination under Title VII because the hiring is not yet complete, and that it is "too late" to challenge the test under the *Beecher* decree. These positions, however, are not inconsistent because the standards of Title VII and the *Beecher* decree are different though related. On the one hand, as the first prima facie step of the Title VII framework, the plaintiffs must show that the examination has caused an adverse and disparate impact. *Steamship Clerks,* 48 F.3d at 601. On the other hand, the *Beecher* decree
177 requires no showing of disparate impact. *177 Instead, the *Beecher* decree requires that the HRD validate the examination. No parties to the *Beecher* decree objected before the HRD put the 2002 and 2004 examinations into use for hiring. Therefore, while the Court finds a *Beecher* decree violation, the objection is untimely for purposes of contempt relief.

ORDER

I order entry of judgment in favor of the plaintiff class regarding liability under Title VII for the 2002 and 2004 entry-level firefighter examinations. The plaintiffs shall propose a remedy within thirty days, and the defendants shall respond within thirty days. In addition, the plaintiffs shall propose a schedule with respect to the entry-level police class as well as the separate allegations involving Lynn. The Court will hold a hearing on *October 30, 2006 at 3 p.m.*

[1] On May 31, 2005, the Court allowed the New England Area Conference of the NAACP and the Boston Society of the Vulcans to intervene. On June 2, 2006, the Court denied the intervenor plaintiffs' request for preliminary relief to require the defendants to reorder the certification list based on the 2004 civil service examination, which continues to be used for the hiring of firefighters in Boston. (Docket No. 119.)

[2] On December 6, 2005, the Court ruled that the HRD constitutes a Title VII employer with respect to the entry-level hiring of firefighters and police officers. (Docket No. 60.)

[3] The HRD designed the 2002 and 2004 examinations to test nine cognitive ability areas: memorization, visualization, spatial orientation, verbal comprehension, verbal expression, information ordering, problem sensitivity, deductive reasoning, and inductive reasoning. (Ex. 10, Attach. C, at 12-38; Ex. 12, Attach. B., at 12-38.) The 2004 examination, for example, gave applicants five minutes to study an illustrated scene of a fire and subsequently, asked seventeen questions of memorization, such as "How many motor vehicles are shown in the scene?" and "What is the license plate number of the Engine 3 fire truck?" (Ex. 30.)

[4] "Evidence of the validity of a test or other selection procedure by a criterion-related validity study should consist of empirical data demonstrating that the selection procedure is predictive of or significantly correlated with important elements of job performance." 29 C.F.R. § 1607.5(B).

[5] "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).

[6] Pub.L. No. 102-166, 105 Stat. 1071 (codified in scattered sections of 2, 16 29, 42 U.S.C.) amended Section 703 of the Civil Rights Act of 1964, 42 U.S.C. § 2000e-2.

[7] For the 2002 examination, 878 candidates identified themselves as African American, 835 as Hispanic, and 10,152 as majority. (Ex. 10, at 4.) For the 2004 examination, 458 candidates identified themselves as African American, 387 as Hispanic, and 6185 as nonprotected group members. (Ex. 12, at 5-6.)

[8] The HRD did not provide 2002 examination data distinguishing Hispanics from "other minorities." (Trial Tr. 69-70, Apr. 11, 2006.) The EEOC Guidelines require that Title VII employers maintain records by enumerated race classifications, including Blacks and Hispanics. 29 C.F.R. § 1607.4(B). Where the data has not been maintained as required, "the Federal enforcement agencies may draw an inference of adverse impact of the selection process from the failure, if [there is] an underutilization of a group in the job category, as compared to the group's representation in the relevant labor market." *Id.* § 1607.4(D).

[9] Neither side produced any evidence on differences with respect to the racial populations of disabled veterans versus non-disabled veterans. The primary focus at trial was on the veterans category, which includes both disabled and non-disabled. Thus, that will be my focus as well.

[10] Boston did not hire any candidates from the 2002 examination and had not, by the end of this trial, hired any candidates without veteran status from the 2004 examination. Between the 2002 and 2004 examinations, Boston hired only white firefighters as a result of the *Quinn* case, which resulted in Boston being released from the *Beecher* minority certification quotas. *See Quinn,* 325 F.3d at 18.

[11] The Court examines Dr. Jacobs's "equity analysis" tables (Ex. 33) rather than his expert report (Ex. 26, at 12), because the expert report does not distinguish between veterans and non-veterans.

[12] 29 C.F.R. § 1607.4(D) provides:

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.

[13] The parties and their experts engaged in a back-and-forth statistical battle during and after trial that added at least five supplemental responses to the record, four of which were submitted after the last day of testimony on May 4, 2006 and two of which arrived after oral argument on June 9, 2006. (*See* Exs. 33R, 33S, 33T, 33IJ; Docket No. 125.) While the Court offered, neither party requested that the record be reopened so that the experts could testify and be cross-examined. Because this factor was less well vetted, I give it less weight.

[14] The plaintiffs attack these reports on other grounds. For example, the plaintiffs assert that the HRD has produced none of the appendices to the reports containing the underlying data.

[15] While no parties pressed this point, Title VII expressly prohibits employers from "adjust[ing] the scores of . . . or otherwise alter[ing] the results of, employment related tests on the basis of race." 42 U.S.C. § 2000e-2(*l*).

[16] Several jurisdictions use personality tests, including Baltimore, Chicago, and Minneapolis. (Exs. 33N, 33*O*.) Personality tests are a fairly new development, and experts have concerns about applicants faking. Jurisdictions such as New Jersey thus have implemented personality tests with appropriate skepticism. (Ex. 25.) The HRD has just developed a new test that includes this component.

[17] As explained by Dr. Jacobs, biodata tests include questions about a candidate's background, such as: "Did you belong to a lot of clubs when you were in high school?" or "Do other people consider you to be motivated?" "There aren't right or wrong answers. There are certain answers that garner more points. I can't tell you. It's trade secret." (Trial Tr. 82, Apr. 14, 2006.)