UNITED STATES DISTRICT COURT
EASTERN DISTRICT OF NEW YORK
----------X
UNITED STATES OF AMERICA,

Plaintiff,

-and-

THE VULCAN SOCIETY, INC., MARCUS HAYWOOD, CANDIDO NUÑEZ, ROGER GREGG,

Plaintiff-Intervenors,

-against-

THE CITY OF NEW YORK, FIRE DEPARTMENT OF THE CITY OF NEW YORK, NEW YORK CITY DEPARTMENT OF CITYWIDE ADMINISTRATIVE SERVICES, and MAYOR MICHAEL BLOOMBERG and NEW YORK CITY FIRE COMMISSIONER NICHOLAS SCOPPETTA, in their individual and official capacities,

Defendants.
----------X

MEMORANDUM & ORDER
07-cv-2067 (NGG)(RLM)

NICHOLAS G. GARAUFIS, United States District Judge.

From 1999 to 2007, the New York City Fire Department used written examinations with discriminatory effects and little relationship to the job of a firefighter to select more than 5,300 candidates for admission to the New York City Fire Academy. These examinations unfairly excluded hundreds of qualified people of color from the opportunity to serve as New York City firefighters. Today, the court holds that New York City's reliance on these examinations constitutes employment discrimination in violation of Title VII of the Civil Rights Act of 1964.

I. INTRODUCTION

In recent years, black and Hispanic residents of New York City (the "City") have come to comprise a substantial portion of the City's population, but their representation in the New York City Fire Department ("FDNY") has remained extraordinarily low.[1] In 2002, the New York City Department of City Planning identified 25% of the City's residents as black and 27% of its residents as Hispanic.[2] At the same time, however, only 2.6% of its firefighters were black and 3.7% of its firefighters were Hispanic.[3] When this litigation commenced in 2007, the percentages of black and Hispanic firefighters had increased to just 3.4% and 6.7%, respectively.[4] In other words, on a force of 8,998 firefighters, there were just 303 black firefighters and 605 Hispanic firefighters. These numbers stand in stark contrast to some of the nation's other large cities, such as Los Angeles, Chicago, Philadelphia, and Houston, where minority firefighters have been represented in significantly higher percentages.[5]

In this case, Plaintiff United States of America (the "Federal Government") as well as the Vulcan Society, Inc., Marcus Haywood, Candido Nuñez and Roger Gregg (the "Intervenors"), have sued to enforce the right of black and Hispanic candidates to be treated fairly in the

[1] The parties use the terms "black" and "Hispanic" in their submissions, and the court adopts the parties' terminology.

[2] See Declaration of Sharon Seeley dated January 21, 2009 (Docket Entry # 253) app. B (citing statistics from the New York City Department of City Planning).

[3] See id.

[4] See id. app. C.

[5] See Declaration of Richard A. Levy dated February 2, 2009 (Docket Entry # 264) Ex. D (showing percentages of minority firefighter representation in 1999 in New York City as 2.9% African-American and 2.8% Hispanic, in Los Angeles as 14.0% African-American and 30.0% Hispanic, in Chicago as 20.4% African-American and 8.6% Hispanic, in Houston as 17.1% African-American and 13.9% Hispanic, and in Philadelphia as 26.3% African-American and 3.2% Hispanic). The 2000 Census figures for those cities show that Los Angeles had 11.2% black residents and 46.5% Hispanic or Latino residents, Chicago had 36.8% black residents and 26.0% Hispanic or Latino residents, Houston had 25.3% black residents and 37.4% Hispanic or Latino residents, and Philadelphia had 43.2% black residents and 8.5% Hispanic or Latino residents. See U.S. Census Bureau, State & County QuickFacts, available at http://quickfacts.census.gov/qfd/index.html (last visited on July 21, 2009).

application process for positions in the FDNY. Specifically, the Federal Government and the Intervenors ("Plaintiffs")[6] challenge the City's reliance on two written examinations that are used to appoint entry-level firefighters to classes at the New York City Fire Academy ("Academy"). These examinations—Written Examination 7029 and Written Examination 2043—were administered from 1999 to 2007, and the City has appointed more than 5,300 entry-level firefighters based upon their results. Although Plaintiffs identify approximately 3,100 of the examination candidates as black and approximately 4,200 of the examination candidates as Hispanic, the City has appointed just 184 black firefighters and 461 Hispanic firefighters from the challenged examinations. (See Section III.A, infra.)

Plaintiffs assert that the City's reliance on Exams 7029 and 2043 in selecting entry-level firefighters has had a disparate impact on black and Hispanic candidates in violation of Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e, et seq. ("Title VII"). The Intervenors also claim, under a disparate treatment theory, that the City, two city agencies, the Mayor and the Fire Commissioner "have long been aware of the discriminatory impact on blacks of their examination process," and that their "continued reliance on and perpetuation of these racially discriminatory hiring processes constitute intentional race discrimination . . . ." (Intervenors' Compl. (Docket Entry # 48) ¶ 51.)

To remedy these claimed violations, Plaintiffs seek various forms of injunctive and monetary relief. The Federal Government seeks to enjoin the City from engaging in discriminatory practices "against blacks on the basis of race and against Hispanics on the basis of

[6] The Intervenors present claims relating only to black firefighters, while the Federal Government brings claims relating to both black and Hispanic firefighters. Although their Motions are separate, the fundamental issues in each are the same. The court will, therefore, generally refer to both parties as "Plaintiff" throughout, distinguishing between the Federal Government and the Intervenors as needed.

national origin," and seeks a specific injunction against the practices challenged in this case. (See Compl. (Docket Entry # 1) ¶ 38.) It also asks the court to order the City to take "appropriate action to correct the present effects of its discriminatory policies and practices" and to enjoin it from failing to "make whole" those harmed by the City's policies and practices. (Id.) The Intervenors seek similar, but broader relief, including an injunction requiring the City to "appoint entry-level firefighters from among qualified black applicants in sufficient numbers to offset the historic pattern and practice of discrimination against blacks in testing and appointment to that position." (Int. Compl., Prayer For Relief ¶ 3(d).) The Intervenors seek to require the City to "recruit black candidates and implement and improve long-range recruitment programs" and to "provide . . . future test scores, appointment criteria, eligibility lists, appointment data, and all other information necessary to conduct an adverse impact and job-relatedness analysis of the examination and selection process." (Id. ¶¶ 3(e), (f).) The Intervenors also seek damages and other fees. (Id. ¶¶ 4-9.)

This is not the first time the City has been brought to federal court to defend its entry-level firefighter examinations against charges of discrimination. In the early 1970s, Judge Weinfeld in the Southern District of New York found that the City's written and physical examinations for entry-level firefighters violated the Equal Protection Clause of the Constitution because of their discriminatory impact on black and Hispanic applicants. See Vulcan Soc'y of New York City Fire Dep't, Inc. v. Civil Serv. Comm'n, 360 F. Supp. 1265, 1269 (S.D.N.Y. 1973), affirmed in relevant part by 490 F.2d 387 (2d Cir. 1973). Following Judge Weinfeld's decision in Vulcan Society, the City contracted with a private consulting firm to construct valid written and physical examinations; these contracts were cancelled three years later, however,

apparently on account of a budget crisis. See Berkman v. City of New York, 536 F. Supp. 177, 184 (E.D.N.Y. 1982).[7] At the time of Vulcan Society, Judge Weinfeld noted the "overwhelming disparity between minority representation in the [FDNY] (5%) and in the general population of New York City within the age group eligible for appointment (32%)." Vulcan Soc'y, 360 F. Supp. at 1269. In the three decades that have followed, these minority groups have come to represent an even greater share of the City's population. Despite these changes, the overwhelmingly monochromatic composition of the FDNY has stubbornly persisted.[8]

This court has already issued several decisions in the case. I have bifurcated the liability and relief phases (see Docket Entry # 47), permitted intervention by the Intervenors (see id.), denied the Intervenors' motion to amend the Intervenors' Complaint (see Docket Entry # 182), declined to dismiss the Intervenors' Complaint on timeliness grounds (see Docket Entry # 231), and certified a class consisting of black applicants for the position of entry-level firefighter (see Docket Entry # 281).

Now before the court are Motions for Summary Judgment by the Federal Government and the Intervenors. (Docket Entries ## 251, 260.) The Federal Government and the Intervenors have moved for summary judgment on the prima facie case of disparate impact, and the Federal

[7] In Berkman, females candidates for the job of entry-level firefighter challenged the physical examination component of a prior FDNY test. See Berkman, 536 F. Supp. at 205-16.

[8] By contrast, it has been reported that the composition of the New York City Police Department ("NYPD") has undergone significant diversification since the 1970s. In 1978, the NYPD was made up of 8.9% black officers and 3.8% Hispanic officers. See Guardians Ass'n of New York City Police Dep't, Inc. v. Civil Serv. Comm'n, 484 F. Supp. 785, 793 (S.D.N.Y. 1980). According to a New York Times article published in June 2009, the NYPD reported its rank-and-file officers as 18% black and 28.7% Hispanic. See Al Baker, "Police Commissioner Plans to Put More Minority Officers in Top Posts," N.Y. Times A20 (June 26, 2009), available at http://www.nytimes.com/2009/06/26/nyregion/26nypd.html (last visited on July 21, 2009). On July 15, 2009, the NYPD announced that it had sworn in its most diverse Police Academy class ever, including 14.7% black recruits and 33.3% Hispanic recruits. See New York City Police Dep't, Press Release # 024, dated July 15, 2009, available at http://www.nyc.gov/html/nypd/html/pr/pr_2009_024.shtml (last visited on July 21, 2009).

Government has joined the Intervenors' Motion for Summary Judgment on the City's business necessity defense.[9] (See Docket Entries ## 251, 260, 263.)

Upon consideration of the parties' submissions and oral argument, the court concludes that Plaintiffs have established a prima facie case that the City's use of the two written examinations has resulted in a disparate impact upon black and Hispanic applicants for the position of entry-level firefighter. The court also concludes that the City has failed to present sufficient evidence supporting a business justification for its employment practices. I therefore grant Plaintiffs' Motions for Summary Judgment in their entirety.

In essence, my ruling is premised upon two basic conclusions. First, Plaintiffs have shown that there is no triable issue of fact as to whether the City's use of Written Exams 7029 and 2043 has resulted in a statistically and practically significant adverse impact on black and Hispanic firefighter applicants. Black and Hispanic applicants disproportionately failed the

[9] In support of its Motion, the Federal Government has submitted a Statement of Undisputed Facts Pursuant to Rule 56.1 (Docket Entry # 252) ("USA 56.1"), a Declaration of Sharon A. Seeley (Docket Entry # 253) ("Seeley Decl."), an Affidavit of Bernard R. Siskin, Ph.D. (Docket Entry # 254) ("Siskin Aff."), including a copy of his November 2007 expert report at Exhibit A ("Siskin Report"), a Memorandum of Law (Docket Entry # 255) ("USA Mem."), and a Reply Memorandum (Docket Entry # 258) ("USA Reply"). In response, the City has submitted an opposing Memorandum of Law on the prima facie case (Docket Entry # 256) ("Def. PF Mem.") and a Response to the Federal Government's Rule 56.1 Statement (Docket Entry # 257) ("Def. USA 56.1").

In support of their Motion, the Intervenors have submitted a Rule 56.1 Statement of Undisputed Facts (Docket Entry # 261) ("Int. 56.1"), a Memorandum of Law (Docket Entry # 262) ("Int. Mem."), a Declaration of Richard A. Levy (Docket Entry # 264) ("Levy Decl."), a Reply Memorandum on the prima facie case (Docket Entry # 268) ("Int. PF Reply") and a Reply Memorandum on job-relatedness and business necessity (Docket Entry # 269) ("Int. BN Reply"). Included in the Levy Declaration are excerpts from the expert report of Dr. Wiesen (Levy Decl. Exs. R, Z ("Wiesen Report").) A full version of the Wiesen Report appears as Docket Entry # 123. In response, the City has submitted an opposing Memorandum of Law relating to the prima facie case ("Def. PF Mem."), an opposing Memorandum of Law relating to job-relatedness and business necessity (Docket Entry # 266) ("Def. BN Mem."), a Response to Intervenors' Rule 56.1 Statement (Docket Entry # 265) ("Def. Int. 56.1"), and a Declaration of William S.J. Fraenkel (Docket Entry # 267) ("Fraenkel Decl."). Included in the Fraenkel Declaration is the expert report of Dr. Schemmer and Dr. Bobko. (Fraenkel Decl. Ex. 1 ("Bobko-Schemmer Report").) The Federal Government has submitted a memorandum joining the Intervenors on job-relatedness and business necessity (Docket Entry # 263 ("USA BN Mem."), and the City has filed a response (Docket Entry # 272 ("Def. USA Resp.")).

The Intervenors have also moved to strike two declarations submitted by the City with their summary judgment papers. (See Docket Entries ## 273, 274, 276, 277.) The court addresses this Motion in Section IV.B below.

written examinations, and those who passed were placed disproportionately lower down than white candidates on the hierarchical hiring lists resulting from their scores. Second, although the City has had the opportunity to justify this adverse impact by showing that it used the written examinations to test for the relevant skills and abilities of entry-level firefighters, the City has failed to raise a triable issue on this defense. Under Second Circuit precedent, the evidence presented by the City is insufficient as a matter of law to justify its reliance on the challenged examinations.

Before proceeding to the legal analysis, I offer a brief word about the Supreme Court's recent decision in Ricci v. DeStefano, 129 S. Ct. 2658 (June 29, 2009). I reference Ricci not because the Supreme Court's ruling controls the outcome in this case; to the contrary, I mention Ricci precisely to point out that it does not. In Ricci, the City of New Haven had set aside the results of a promotional examination, and the Supreme Court confronted the narrow issue of whether New Haven could defend a violation of Title VII's disparate treatment provision by asserting that its challenged employment action was an attempt to comply with Title VII's disparate impact provision. The Court held that such a defense is only available when "the employer can demonstrate a strong basis in evidence that, had it not taken the action, it would have been liable under the disparate-impact statute." Id. at 2664. In contrast, this case presents the entirely separate question of whether Plaintiffs have shown that the City's use of Exams 7029 and 2043 has actually had a disparate impact upon black and Hispanic applicants for positions as entry-level firefighters. Ricci did not confront that issue.

The Ricci Court concluded that New Haven would not likely have been liable under a disparate impact theory. See id. at 2681. In doing so, the Court relied on the various steps that

New Haven took to validate its civil service examination. Id. at 2678-79. It is noteworthy, however, that in this case New York City has taken significantly fewer steps than New Haven took in validating its examination. The relevant teaching of Ricci, in this regard, is that the process of designing employment examinations is complex, requiring consultation with experts and careful consideration of accepted testing standards. As discussed below, these requirements are reflected in federal regulations and existing Second Circuit precedent. This legal authority sets forth a simple principle: municipalities must take adequate measures to ensure that their civil service examinations reliably test the relevant knowledge, skills and abilities that will determine which applicants will best perform their specific public duties.

In rendering this decision, I am aware that the use of multiple-choice examinations is typically intended to apply objective standards to employment decisions. Similarly, I recognize that it is natural to assume that the best performers on an employment test must be the best people for the job. But, the significance of these principles is undermined when an examination is not fair. As Congress recognized in enacting Title VII, when an employment test is not adequately related to the job for which it tests—and when the test adversely affects minority groups—we may not fall back on the notion that better test takers make better employees. The City asks the court to do just that. Regrettably, though, the City did not take sufficient measures to ensure that better performers on its examinations would actually be better firefighters. Accordingly, the court grants the Motions for Summary Judgment and finds that Plaintiffs have established disparate impact liability.

II. SUMMARY JUDGMENT STANDARD

"Summary judgment is appropriate when the pleadings and admissible evidence proffered to the district court show that there is 'no genuine issue as to any material fact and that the moving party is entitled to a judgment as a matter of law . . . .'" Major League Baseball Props., Inc. v. Salvino, Inc., 542 F.3d 290, 309 (2d Cir. 2008) (quoting Fed. R. Civ. P. 56(c)). "Material facts are those which 'might affect the outcome of the suit under the governing law,' and a dispute is 'genuine' if 'the evidence is such that a reasonable jury could return a verdict for the nonmoving party.'" Coppola v. Bear Stearns & Co., 499 F.3d 144, 148 (2d Cir. 2007) (quoting Anderson v. Liberty Lobby, Inc., 477 U.S. 242, 248 (1986)). Factual disputes that are irrelevant or immaterial to the disposition of a case cannot preclude a grant of summary judgment. See Loria v. Gorman, 306 F.3d 1271, 1282-83 (2d Cir. 2002).

In considering a motion for summary judgment, the court construes the facts "in the light most favorable to the nonmoving party," and draws "all reasonable inferences in its favor." SCR Joint Venture L.P. v. Warshawsky, 559 F.3d 133, 137 (2d Cir. 2009). "[T]he moving party bears the burden of demonstrating the absence of a genuine issue of material fact." Baisch v. Gallina, 346 F.3d 366, 371-72 (2d Cir. 2003) (citing Celotex Corp. v. Catrett, 477 U.S. 317, 323 (1986)). In response, the nonmoving party "'must do more than simply show that there is some metaphysical doubt as to the material facts . . . .'" Jeffreys v. City of New York, 426 F.3d 549, 554 (2d Cir. 2005) (quoting Matsushita Elec. Indus. Co. v. Zenith Radio Corp., 475 U.S. 574, 586 (1986)).

III. THE PRIMA FACIE CASE

Plaintiffs seek summary judgment on their prima facie case of disparate impact. As set forth below, summary judgment on Plaintiffs' prima facie case is warranted. The facts set forth below are undisputed, unless otherwise noted.

A. The Hiring Process

During the relevant period, the hiring of entry-level firefighters from Exams 7029 and 2043 proceeded in several stages. Candidates interested in a position as an entry-level firefighter began by submitting an application to the Department of Citywide Administrative Services ("DCAS"), paid an application fee (unless it had been waived), and received an admission card for a written examination. (See USA 56.1 ¶¶ 13, 14; Int. 56.1 ¶ 9.) Each written examination was an 85-question, paper-and-pencil multiple choice test, and an applicant's raw score on that examination was simply a percentage of the questions answered correctly. (See USA 56.1 ¶¶ 16, 23; Int. 56.1 ¶ 11.) A passing score was set for each examination, and after the results were in, the City notified each applicant of his or her score, as well as whether he or she had passed. (See Int. 56.1 ¶ 17.) Versions of each examination (with the same questions, but sometimes different question-orderings) were administered on repeated occasions—Exam 7029 was administered from 1999 through 2002, and Exam 2043 was administered from 2002 through 2007. (See USA 56.1 ¶¶ 17-22.)

Candidates who passed the written examination were allowed to take the physical performance test ("PPT"), but those who failed the written examination could not take the PPT. (See id. ¶ 27; Int. 56.1 ¶ 14.) The PPT consisted of eight physical tasks, and a candidate had to pass a minimum of six tasks to achieve a passing score overall. (Int. 56.1 ¶ 19.) A passing

candidate's score on the PPT was simply a percentage of the number of tasks successfully completed. For example, passing eight tasks resulted in a score of 100%, passing seven tasks resulted in a score of 87.5%, and passing six tasks resulted in a score of 75%. (See id.)

Candidates who passed both the written examination and the PPT were placed on a "rank-order" eligibility list. (USA 56.1 ¶¶ 28-30; Seeley Decl. app. I.) The ordering of the eligibility list was based upon an elaborate process of, inter alia, "standardizing," "combining," and "transforming" the raw scores. (See USA 56.1 ¶ 31; Int. 56.1 ¶ 20.) Specifically, the raw score from the written examination and the PPT would be "standardized" by subtracting the average score for all candidates from an individual candidate's score and then dividing that number by the standard deviation for the test. (See Seeley Decl. apps. J, K.) The resulting scores from both the written examination and the PPT would then be divided in half and added together to create a "Combined Weighted Standard Score." (Id.) The Combined Weighted Standard Score was then converted into a "Transformed Score" by multiplying by either 18.472906403940886699 (for Exam 7029) or 12.7226 (for Exam 2043), and then adding either 83.74384236453 (for Exam 7029) or 88.4606 (for Exam 2043). (Id.) Finally, the "Adjusted Final Average," used to rank candidates on the eligibility list, was created by adding any "Residency," "Legacy," or "Veteran" points to the Transformed Score. (Id.; see USA 56.1 ¶¶ 30, 31; Int. 56.1 ¶ 20.) This elaborate process resulted in a list of candidates eligible to be appointed to Academy classes in order of rank.

The DCAS and FDNY would determine how many candidates would be needed to fill an upcoming class and would certify a portion of the eligibility list for appointment, beginning with the highest scores. (USA 56.1 ¶¶ 40-41; Int. 56.1 ¶ 23.) The FDNY's Candidate Investigation

Division ("CID") took steps to process and investigate candidates in order of their ranking on the eligibility list. (USA 56.1 ¶¶ 36-37, 39; Int. 56.1 ¶ 24.) This investigation involved, inter alia, background checks, intake interviews, and medical and psychological examinations by the FDNY's Bureau of Health Services. (USA 56.1 ¶¶ 35-37; Int. 56.1 ¶ 24; Seeley Decl. app. A (Request for Admission # 101).) The City would fill slots in an Academy class by proceeding down the list of eligible and qualified applicants until the class was filled; once a class was filled, any eligible and qualified candidate still remaining would not be appointed for that class. (USA 56.1 ¶¶ 42, 43; Int. 56.1 ¶ 25.)

Written Examination 7029 was first administered on February 26, 1999, and versions of it were administered as late as 2002. (USA 56.1 ¶¶ 17-19; Int. 56.1 ¶ 15.) Approximately 1,750 black applicants and approximately 2,125 Hispanic applicants sat for Exam 7029. (Siskin Report tbls. 1, 2.) The City hired from the eligibility list resulting from Exam 7029 from February 2001 through at least September 2004, and appointed over 3,200 entry-level firefighters from that examination. (USA 56.1 ¶¶ 11, 44.) Of this number, 104 (3.2%) individuals were black and 274 (8.5%) individuals were Hispanic. (Id. ¶ 11.)

Written Examination 2043 was first administered on December 14, 2002, and versions of it were administered as late as March 2007. (Id. ¶¶ 20-22; Int. 56.1 ¶ 16.) Approximately 1,390 black applicants and approximately 2,125 Hispanic applicants sat for Exam 2043. (Siskin Report tbls. 5, 6.) The City hired firefighters from the eligibility list resulting from Exam 2043 from May 2004 through at least January 2008, and, as of November 2007, the City had appointed over 2,100 entry-level firefighters from that examination. (USA 56.1 ¶¶ 12, 46.) Of this number, 80 (3.7%) were black and 187 (8.7%) were Hispanic. (Id. ¶ 12.)

With these undisputed background facts in mind, the court addresses the prima facie case.

B. The Use of Statistics for a Prima Facie Case

A prima facie showing of disparate impact "requires plaintiffs to establish by a preponderance of the evidence that the employer 'uses a particular employment practice that causes a disparate impact on the basis of race, color, religion, sex, or national origin.'" Robinson v. Metro-North Commuter R.R. Co., 267 F.3d 147, 160 (2d Cir. 2001) (quoting 42 U.S.C. § 2000e-2(k)(1)(A)(i)). "To make this showing, a plaintiff must (1) identify a policy or practice, (2) demonstrate that a disparity exists, and (3) establish a causal relationship between the two." Id. at 160.

Statistics alone can make out the prima facie case. See EEOC v. Joint Apprenticeship Comm. of Joint Indus. Bd. of Elec. Indus., 186 F.3d 110, 117 (2d Cir. 1999); see also Robinson, 267 F.3d at 160 ("[S]tatistical proof almost always occupies center stage in a prima facie showing of a disparate impact claim."). In order to do so, "[t]he statistics must reveal that the disparity is substantial or significant." Robinson, 267 F.3d at 160 (internal quotation marks omitted). "Moreover, the statistics must be of a kind and degree sufficient to reveal a causal relationship between the challenged practice and the disparity." Id. "[A] plaintiff may establish a prima facie case of disparate impact discrimination by proffering statistical evidence which reveals a disparity substantial enough to raise an inference of causation. That is, a plaintiff's statistical evidence must reflect a disparity so great that it cannot be accounted for by chance." Joint Apprenticeship Comm., 186 F.3d at 117.

There are at least two widely recognized statistical measures of disparate impact: (1) the 80% or Four-fifths Rule, and (2) statistical significance or standard deviation analysis. See

Atkins v. Westchester County Dep't of Soc. Serv., 31 Fed. Appx. 52, 53 (2d Cir. 2002) (summary order) ("[i]n evaluating disparate impact claims under Title VII, this Court has primarily relied upon [these] two methods of measuring disparities between groups"). Federal regulations set out the 80% Rule, and courts have recognized it as a "rule of thumb" for statistical analysis of disparate impact. See, e.g., Joint Apprenticeship Comm., 186 F.3d at 118 ("This rule is not binding on courts, and is merely a 'rule of thumb' to be considered in appropriate circumstances."); see also United States v. New York City Bd. of Educ., 487 F. Supp. 2d 220, 224 (E.D.N.Y. 2007). The 80% Rule appears at 29 C.F.R. § 1607.4D, which states:

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

Id. Essentially, this means that if the minority group performs less than 80% as well as the highest performing group, disparate impact will generally be inferred.

Courts have also relied upon standard deviation analysis (or statistical significance analysis) in determining whether there has been a disparate impact. "Standard deviation analysis measures the probability that a result is a random deviation from the predicted result—the more standard deviations the lower the probability the result is a random one." Waisome v. Port Auth. of New York & New Jersey, 948 F.2d 1370, 1376 (2d Cir. 1991); see also Barbara Lindemann & Paul Grossman, Employment Discrimination Law ("Lindemann") 94 (3d ed. 1996) ("Tests of statistical significance are commonly used in the social sciences to rule out chance as the cause

of observed disparities."). "Basically, looking at standard deviations indicates how far an obtained result varies from an expected result." Smith v. Xerox Corp., 196 F.3d 358, 365 (2d Cir. 1999), overruled on unrelated grounds by Meacham v. Knolls Atomic Power Lab., 461 F.3d 134, 141 (2d Cir. 2006). The Second Circuit has "looked to whether the plaintiff can show a statistically significant disparity of two standard deviations" in making a prima facie showing. Id. at 365.

"Although courts have considered both the four-fifths rule and standard deviation calculations in deciding whether a disparity is sufficiently substantial to establish a prima facie case of disparate impact, there is no one test that always answers the question." Id. at 366. According to either measurement, "the substantiality of a disparity is judged on a case-by-case basis." Id.

C. The Statistics for Plaintiffs' Prima Facie Case

Plaintiffs allege that four employment practices related to the challenged examinations have had an unlawful disparate impact on black and Hispanic candidates for the position of entry-level firefighter. Specifically, they challenge the City's use of:

(1) Written Examination 7029 as a pass/fail screening device with a cutoff score of 84.705;

(2) Rank-order processing and selection of candidates from the Written Examination 7029 eligibility list based on a combination of their scores on Written Examination 7029 and the PPT;

(3) Written Examination 2043 as a pass/fail screening device with a cutoff score of 70;

(4) Rank-order processing and selection of candidates from the Written Examination 2043 eligibility list based on a combination of their scores on Written Examination 2043 and the PPT.

(USA Mem. 1; see Int. Mem. 7.) As this court previously noted, "[t]he use of an examination as a 'pass/fail screening device' means the use [of] the examination to exclude from appointment those applicants that have failed the examination." (Memorandum & Order dated May 11, 2009 (Docket Entry # 281), at 3.) "The use of an examination as a part of 'rank-order processing' means the use of the examination as a component of the overall score that determines an applicant's position on a hierarchical hiring list." (Id.)

According to the Federal Government, the statistical significance analysis performed by its expert, Bernard R. Siskin, Ph.D., establishes a prima facie case of disparate impact. (USA Mem. 2.) Similarly, the Intervenors argue that the statistical significance analyses performed by Dr. Siskin and their expert, Joel P. Wiesen, Ph.D., establish a prima facie case. (Int. Mem. 7.)

The court briefly reviews the statistics that Plaintiffs have presented on each of the challenged practices. The City does not dispute the statistical calculations of Plaintiffs' experts, but rather, disputes Plaintiffs' reliance on statistical significance testing because of assumptions underlying that methodology. The court will address these assumptions after setting out the undisputed statistical calculations of Plaintiffs' experts.

1. Pass/Fail Use of Exam 7029

The cutoff passing score for Written Examination 7029 was 84.705%. (USA 56.1 ¶ 24; Int. 56.1 ¶ 13.) Based on that cutoff score, the pass rate of white candidates for Exam 7029 was 89.9%, while the pass rate of black candidates was 60.3%. (USA 56.1 ¶ 83; Int. 56.1 ¶ 28.) In other words, out of 12,915 white test takers, 11,613 received a passing score of at least 84.705, whereas out of 1,749 black test takers, only 1,054 received a passing score. (See Int. 56.1 ¶ 27; Wiesen Report, tbl. 3a.) The pass rate of black candidates was, therefore, 67% of the pass rate of

white candidates. (USA 56.1 ¶ 89.) Both Dr. Siskin's and Dr. Wiesen's standard deviation analysis found that this disparity is equivalent to 33.9 units of standard deviation, meaning that the likelihood it occurred by chance is less than 1 in 4.5 million-billion. (See id. ¶ 85 (citing, inter alia, Siskin Report 3, 21); see also Int. 56.1 ¶ 30 (citing, inter alia, Wiesen Report 18-19).)

The practical effect of this disparity, according to Dr. Siskin, is that 519 black candidates who failed the examination—74.7% of the black applicants who failed—were eliminated from consideration. (See USA 56.1 ¶¶ 86-87 (citing Siskin Report).) Dr. Wiesen estimated that 457 black candidates would have passed the examination but for the effect of this disparity. (See Int. 56.1 ¶ 38 (citing Wiesen Report).) Based on Dr. Siskin's calculation, 114 additional black firefighters would have been appointed absent the disparity. (USA 56.1 ¶ 88 (citing Siskin Report).) This last calculation was based on the assumption that the black applicants who failed Exam 7029 would have passed the PPT at the same rate as other similarly situated passers, and would have met the other qualifications and been appointed at the same rate as other passers.[10] (Siskin Report 16-17.)

The pass rate for Hispanic candidates taking Exam 7029 was 76.7%, compared with a pass rate of 89.9% for white candidates. (USA 56.1 ¶ 92.) Accordingly, the pass rate of Hispanic candidates was 85.3% of the pass rate of white candidates. (Siskin Report, tbl. 2.) Dr. Siskin's standard deviation analysis found that this disparity is equivalent to 17.4 units of standard deviation, meaning that the likelihood it occurred by chance is less than 1 in 4.5 million-billion. (USA 56.1 ¶ 94 (citing, inter alia, Siskin Report 3, 23).)

[10] The City accepts as undisputed Dr. Siskin's calculation concerning those who would have been appointed absent the disparity, assuming "use of a test of statistical significance, the assumption of 'shortfall' suggested by Dr. Siskin and inferring no difference in the capability and preparedness of the groups being compared." (Def. USA 56.1 ¶ 88; see also id. ¶¶ 86-88, 95-97, 103-05, 111-13.) As discussed below, to the extent that the City challenges these assumptions, the court concludes that its arguments are without merit and are insufficient to preclude summary judgment.

The practical effect of this deviation, according to Dr. Siskin, is that 282 Hispanic candidates who failed the examination—56.9% of the Hispanic applicants who failed—were eliminated from consideration. (See id. ¶¶ 95-96 (citing Siskin Report).) Based on Dr. Siskin's calculation, 62 additional Hispanic firefighters would have been appointed absent the disparity. (Id. ¶ 97 (citing Siskin Report).) This last calculation was based on the assumption that the Hispanic applicants who failed Exam 7029 would have passed the PPT at the same rate as other similarly situated passers, and would have met the other qualifications and been appointed at the same rate as other passers. (Siskin Report 16-17.)

2. Pass/Fail Use of Exam 2043

The cutoff passing score for Written Examination 2043 was 70%. (USA 56.1 ¶ 25; Int. 56.1 ¶ 13.) Based on this cutoff score, the pass rate of white candidates taking Exam 2043 was 97.2%, while the pass rate of black candidates was 85.4%. (USA 56.1 ¶ 100; see Wiesen Report 42.) In other words, out of 13,877 white test takers, 13,495 received a passing score of at least 70, whereas, out of 1,393 black test takers, 1,190 received a passing score. (See Int. 56.1 ¶ 32; Wiesen Report, tbl. 16a.) The pass rate of black candidates was, therefore, 87.8% of the pass rate of white candidates. (Siskin Report, tbl. 5.) Both Dr. Siskin's and Dr. Wiesen's standard deviation analysis found that this disparity is equivalent to 21.8 units of standard deviation, meaning that the likelihood that it occurred by chance is less than 1 in 4.5 million-billion. (See USA 56.1 ¶ 102 (citing, inter alia, Siskin Report 5, 26); see also Int. 56.1 ¶ 35 (citing, inter alia, Wiesen Report 42-43).)

The practical effect of this deviation, according to Dr. Siskin, is that 165 black candidates who failed the examination—81.3% of the black applicants who failed—were eliminated from

consideration. (See USA 56.1 ¶¶ 103-04 (citing Siskin Report).) Dr. Wiesen estimated that 150 black candidates would have passed the examination absent the disparity. (See Int. 56.1 ¶ 39 (citing Wiesen Report).) Based on Dr. Siskin's calculation, 30 additional black firefighters would have been appointed absent the disparity. (USA 56.1 ¶ 105 (citing Siskin Report).) This last calculation was based on the assumption that the black applicants who failed Exam 2043 would have passed the PPT at the same rate as other similarly situated passers, and would have met the other qualifications and been appointed at the same rate as other passers. (Siskin Report 16-17.)

The pass rate for Hispanic candidates taking Exam 2043 was 92.8%, compared with a pass rate of 97.2% for white candidates. (USA 56.1 ¶ 108.) The pass rate of Hispanic candidates was, therefore, 95.5% of the pass rate of white candidates. (Siskin Report, tbl. 6.) Dr. Siskin's standard deviation analysis found that this disparity is equivalent to 10.5 units of standard deviation, meaning that the likelihood it occurred by chance is less than 1 in 4.5 million-billion. (USA 56.1 ¶ 110 (citing, inter alia, Siskin Report 5, 27).)

The practical effect of this deviation, according to Dr. Siskin, is that 94 Hispanic candidates who failed the examination—61.8% of the Hispanic applicants who failed—were eliminated from consideration. (See id. ¶¶ 111-12 (citing Siskin Report).) Based on Dr. Siskin's calculation, 17 additional Hispanic firefighters would have been appointed absent the disparity. (Id. ¶ 113 (citing Siskin Report).) This last calculation was based on the assumption that the Hispanic applicants who failed Exam 2043 would have passed the PPT at the same rate as other similarly situated passers, and would have met the other qualifications and been appointed at the same rate as other passers. (Siskin Report 16-17.)

3. Rank-Ordering Use of Exam 7029

On the eligibility list created from Written Exam 7029 and the PPT, black candidates were grouped disproportionately lower down than white candidates. For example, only 10.1% of black candidates were in the top 20% of all candidates, while 53.8% of black candidates were in the bottom 40%, and 29.2% of black candidates were in the bottom 20%.[11] (USA 56.1 ¶ 120.) While 33% of white candidates had eligibility list numbers at or above 2000, only 21% of black candidates did, and while 20% of white candidates had list numbers at or below 5001, 30% of black candidates did. (Id. ¶¶ 118-19.) According to Dr. Wiesen's calculations, the average ranking of a black candidate was 630 ranking places lower than that of a white candidate, amounting to a disparity of 6.5 units of standard deviation. (See Wiesen Report, tbl. 9b.) According to Dr. Siskin's calculation, the disparity between the placement of black and white candidates on the eligibility list is equivalent to 6.5 units of standard deviation, meaning that the likelihood it occurred by chance is less than 1 in 11 billion. (Siskin Report 24, 25; USA 56.1 ¶ 117.) Dr. Siskin calculated that, on account of this disparity, 68 out of 104 black candidates were delayed in appointment for an aggregate total of approximately 20 years of delayed wages and seniority. (USA 56.1 ¶ 122; Siskin Report, tbl. 3b.)

Similarly, the eligibility list created from the Exam 7029 results placed Hispanic candidates disproportionately lower down than white candidates. Only 14.3% of Hispanic candidates were in the top 20% of all candidates, while 47.8% of Hispanic candidates were in the

[11] The document containing these figures is an email from a DCAS employee (the "DCAS Email"), and suggests that this lower grouping was more pronounced for the written examination. It shows that only 8.7% of black applicants were among the highest 24.8% of scores on Written Exam 7029, but that 44.3% of black applicants were among the bottom 19.7% of scores. (See Seeley Decl. app. AC.) That is, black applicants had about a third as many of the highest scores and more than twice as many of the lowest scores as would be expected absent a disparity.

bottom 40%, and 27.3% of Hispanic candidates were in the bottom 20%.[12] (USA 56.1 ¶ 130.) While 33% of white candidates had eligibility list numbers at or above 2000, only 28% of Hispanic candidates did, and while 20% of white candidates had list numbers at or below 5001, 29% of Hispanic candidates did. (Id. ¶¶ 128-29.) According to Dr. Siskin's calculation, the disparity between the placement of Hispanic and white candidates on the eligibility list is equivalent to 4.6 units of standard deviation, meaning that the likelihood it occurred by chance is less than 1 in 204,000. (Id. ¶ 127.) Dr. Siskin calculated that, on account of this disparity, 86 out of 274 Hispanic candidates were delayed in appointment for an aggregate total of approximately 23 years of delayed wages and seniority. (Id. ¶ 132; Siskin Report, tbl. 4b.)

4. Rank-Ordering Use of Exam 2043

On the eligibility list created from Written Exam 2043 and the PPT, black candidates were grouped disproportionately lower down than white candidates. For example, only 11.4% of black candidates were in the top 20% of all candidates, while 56.9% of black candidates were in the bottom 40%, and 46.2% of black candidates were in the bottom 20%.[13] (USA 56.1 ¶ 139.) While 28% of white candidates had eligibility list numbers at or above 2000, only 18% of black candidates did, and while 30% of white candidates had list numbers at or below 5001, 50% of black candidates did. (Id. ¶¶ 137-38.) According to Dr. Wiesen's calculations, the average ranking of a black candidate was 974 ranking places lower than that of a white candidate,

[12] The DCAS Email shows that only 13.8% of Hispanic applicants were among the highest 24.8% of scores on Written Exam 7029, but that 36.3% of Hispanic applicants were among the bottom 19.7% of scores. (See Seeley Decl. app. AC.) That is, Hispanic applicants had about half as many of the highest scores and about twice as many of the lowest scores as would be expected absent a disparity.

[13] The DCAS Email shows that only 5.3% of black applicants were among the highest 18.1% of scores on Written Exam 2043, but that 48.9% of black applicants were among the bottom 22.3% of scores. (See Seeley Decl. app. AC.) That is, black applicants had less than a third as many of the highest scores and more than twice as many of the lowest scores as would be expected absent a disparity.

amounting to a disparity of 9.6 units of standard deviation. (See Wiesen Report, tbl. 22b.) According to Dr. Siskin's calculation, the disparity between the placement of black and white candidates on the eligibility list was equivalent to 9.5 units of standard deviation, meaning that the likelihood it occurred by chance is less than 1 in 4.5 million-billion. (USA 56.1 ¶ 136.) Dr. Siskin calculated that, on account of this disparity, 44 out of 80 black candidates were delayed in appointment for an aggregate total of approximately fourteen years of delayed wages and seniority. (Id. ¶ 146; Siskin Report, tbl. 12b.)

Similarly, the eligibility list created from the results of Exam 2043 placed Hispanic candidates disproportionately lower down than white candidates. Only 17.2% of Hispanic candidates were in the top 20% of all candidates, while 45.4% of Hispanic candidates were in the bottom 40%, and 24.6% of Hispanic candidates were in the bottom 20%.[14] (USA 56.1 ¶ 152.) While 28% of white candidates had eligibility list numbers at or above 2000, only 25% of Hispanic candidates did, and while 30% of white candidates had list numbers at or below 5001, 39% of Hispanic candidates did. (Id. ¶¶ 150-51.) According to Dr. Siskin's calculation, the disparity between the placement of Hispanic and white candidates on the eligibility list is equivalent to 4.6 units of standard deviation, meaning that the likelihood it occurred by chance is less than 1 in 186,225. (Id. ¶ 149.) Dr. Siskin calculated that, on account of this disparity, 51 out of 187 Hispanic candidates were delayed in appointment for an aggregate total of approximately twelve years of delayed wages and seniority. (Id. ¶ 158; Siskin Report, tbl. 14b.)

[14] The DCAS Email shows that only 10.1% of Hispanic applicants were among the highest 18.1% of scores on Written Exam 2043, but that 34.9% of Hispanic applicants were among the bottom 22.3% of scores. (See Seeley Decl. app. AC.) That is, Hispanic applicants had about half as many of the highest scores and about one-and-a-half times as many of the lowest scores as would be expected absent a disparity.

Dr. Siskin also conducted tests addressing the fact that the eligibility list from Exam 2043 was not exhausted, and, therefore, candidates very low down on that list were never reached. Dr. Siskin determined that those who were never reached "effectively failed" the examination process. (See USA 56.1 ¶ 160.) According to Dr. Siskin's calculations, out of 95 black candidates and 63 Hispanic candidates who would have ranked high enough to be considered for hire, 42 of the black candidates and 28 of the Hispanic candidates would have been appointed absent a disparity resulting from Written Exam 2043. (Siskin Report 33-35; USA 56.1 ¶¶ 145, 157.) Based on the hiring rates of candidates from the Exam 2043 eligibility list, he calculated that the disparity of hiring rates between white and black candidates amounted to 9.7 units of standard deviation, while the disparity of hiring rates between white and Hispanic candidates amounted to 5 units of standard deviation. (USA 56.1 ¶¶ 142, 155.)

Dr. Siskin utilized the data relating those who effectively failed in order to calculate an "effective pass rate" for Written Examination 2043, determined to be 70.3% for white candidates, 41.5% for black candidates, and 58.9% for Hispanic candidates. (Id. ¶¶ 162-63, 167.) This amounts to a statistical disparity between white and black candidates of 21.9 units of standard deviation, and a statistical disparity between white and Hispanic candidates of 10.5 units of standard deviation. (Id. ¶¶ 164, 168.)

D. The Parties' Respective Positions

Plaintiffs argue that the presented statistics establish a prima facie case of disparate impact for the four challenged employment practices. (See USA Mem. 10-17; Int. Mem. 6-12.) They point out that the calculated disparities between black and minority candidates resulting from the challenged practices are much greater than three units of standard deviation. They also

emphasize the practical significance of these disparities—for example, the Federal Government relies on the statistical analyses showing that, but for the disparities resulting from the written examinations, "1,060 additional black and Hispanic candidates would have been considered for appointment as FDNY firefighters," "an estimated 293 additional black and Hispanic candidates would have been appointed as FDNY firefighters," and "249 black and Hispanic firefighters who were appointed—about 39% of those appointed from the examinations at issue in this case—would have been appointed earlier." (USA Mem. 2-3.) Finally, Plaintiffs argue that the City has conceded these statistical conclusions. (See, e.g., id. at 3.)

In opposition to the Motions, the City offers several iterations of the same basic argument. In essence, the City asks the court to reject Plaintiffs' statistical significance analysis because it improperly assumes "perfect parity" among groups of people (see Def. PF Mem. 1-3, 5-7), and erroneously produces a finding of disparate impact solely on account of large sample sizes (see id. at 1, 5, 6, 7). The City asks the court to rely exclusively upon the 80% Rule in determining whether there has been a disparate impact between white and minority candidates. (See id. at 1, 2-3.) Because application of this statistical rule would result in a finding of disparate impact for some, but not all, of the challenged employment practices, the City asks the court to deny summary judgment relating to those practices that do not meet the 80% Rule. (See id. at 7-8.) The City does not contest the specific calculations in Plaintiffs' Rule 56.1 Statements, instead attacking the assumptions on which they rely, and denying the "materiality" of the facts presented.

Plaintiffs respond that large sample sizes do not undermine the validity of statistical significance testing; rather, they argue, it is small sample sizes that render statistical significance

tests less reliable. (See USA Mem. 18-20; Int. Mem. 10.) Plaintiffs also argue that there is no basis for relying on the 80% Rule to the exclusion of statistical significance testing, and that, in fact, all legal authority is to the contrary. (USA Mem. 20-22; USA Reply 3-7; Int. Mem. 13 n.10; Int. PF Reply 2-4.)

Before addressing the parties' respective positions, the court notes that the dispute regarding the proper statistical measurement for disparate impact does not relate to all of the challenged employment practices. It is undisputed that the City's pass/fail use of Exam 7029 has had a disparate impact upon black candidates under both statistical significance testing and under the 80% Rule. (See USA 56.1 ¶¶ 59, 83, 89.) Moreover, the City offers the 80% Rule only as a means of comparing pass rates, not rank-ordering. (See USA 56.1 ¶¶ 67, 69.) It has not presented an alternative statistical measure for the rank-ordering of candidates. (See Def. PF Mem. 7-8.) The City's preference for the 80% Rule, therefore, solely relates to the pass/fail uses of Exam 2043 with respect to black candidates, and the pass/fail uses of Exam 7029 and 2043 with respect to Hispanic candidates.

E. Plaintiffs Have Demonstrated a Prima Facie Case

Plaintiffs have demonstrated a prima facie case of disparate impact by (1) identifying four specific employment practices (each relating to both black and Hispanic applicants), (2) demonstrating that a disparity exists among groups, and (3) establishing a causal relationship between the employment practices and the disparities. See Robinson, 267 F.3d at 160. For each employment practice, Plaintiffs have presented analyses from two experts that thoroughly demonstrate the statistical significance of the disparities between groups of candidates. For each of the pass/fail uses of the examinations, these analyses demonstrate that the disparities between

the pass rates of whites and minority candidates were between 10.5 and 33.9 units of standard deviation. For each of the rank-ordering uses of the examinations, the analyses demonstrate that the disparities between the rankings of whites and minority candidates were between 4.6 and 9.7 units of standard deviation. These statistical disparities show that black and Hispanic candidates disproportionately failed Written Exams 7029 and 2043, and were placed disproportionately lower on the eligibility lists created from those examinations.

The Second Circuit has repeatedly recognized that standard deviations of more than 2 or 3 units can give rise to a prima facie case of disparate impact because of the low likelihood that such disparities have resulted from chance. See Malave v. Potter, 320 F.3d 321, 327 (2d Cir. 2003) ("courts 'generally consider this level of significance [i.e,. two standard deviations] sufficient to warrant an inference of discrimination.'") (quoting Smith, 196 F.3d at 365); Waisome, 948 F.2d at 1376 ("[a] finding of two or three standard deviations (one in 384 chance the result is random) is generally highly probative of discriminatory treatment"); Ottaviani v. State Univ. of New York, 875 F.2d 365, 372 (2d Cir. 1989) ("It is certainly true that a finding of two to three standard deviations can be highly probative of discriminatory treatment."); Guardians Assoc. of New York City Police Dep't, Inc. v. Civil Serv. Comm., 630 F.2d 79, 86 (2d Cir. 1980) ("Guardians") ("[I]n cases involving large samples, 'if the difference between the expected value (from a random selection) and the observed number is greater than two or three standard deviations,' a prima facie case is established.") (quoting Castaneda v. Partida, 430 U.S. 482, 496 n.17 (1977). The calculated standard deviations in this case are all well beyond 2 to 3 units, strongly supporting a conclusion of a causal relationship between the observed disparities and the employment practices at issue.

The significance of Plaintiffs' statistics is bolstered by evidence that the disparities have been significant as a practical matter. See Lindemann 94 ("To guard against the possibility that a finding of adverse impact could result from the statistical significance of a trivial disparity or meaningless difference in results, the Uniform Guidelines on Employee Selection Procedures[, 29 C.F.R. § 1607.4D,] and the courts have adopted an additional test for adverse impact: that a statistically significant disparity also has practical significance."). As mentioned above, approximately one thousand additional black and Hispanic candidates would have been considered for appointment as FDNY firefighters had it not been for the disparities resulting from the examinations. Further, absent these disparities, approximately 293 additional black and Hispanic candidates would have been appointed from the eligibility lists used from 2001 through 2008, and approximately 249 black and Hispanic applicants who were actually appointed would have been appointed sooner. Given that, in 2007, the FDNY had 8,998 firefighters, including only 303 black firefighters and 605 Hispanic firefighters (see Seeley Decl. app. C), it is clear that these disparities have a substantial practical significance. In fact, the disparities are overwhelming.

The accuracy of Plaintiffs' statistical calculations is not disputed, and the City's Responses to Plaintiffs' 56.1 Statements essentially concede the statistical picture establishing a prima facie case. The City specifically concedes that: (1) the disparity created by each of the challenged practices is more than three units of standard deviation (see Def USA 56.1 ¶¶ 84, 93, 101, 109, 116, 126, 135, 148; Def. Int. 56.1 ¶¶ 30, 31, 35, 37), (2) the Plaintiffs' calculations of statistical significance are "undisputed" (see, e.g., Def. USA 56.1 ¶¶ 85, 94, 102, 110, 117, 127, 136, 149), and (3) "[o]ne of the City's experts conducted analyses to attempt to verify Dr.

Siskin's statistical calculations and confirmed the results reported by Dr. Siskin" (USA 56.1 ¶ 77; Def. USA 56.1 ¶ 77). Moreover, the City has essentially admitted the calculations performed by Plaintiffs' experts showing the disparities' practical significance.[15] (See Def. USA 56.1 ¶¶ 86-88, 95-97, 103-105, 111-13, 122, 132, 145, 157; Def. Int. 56.1 ¶¶ 37-39; see also Def. USA 56.1 ¶¶ 82, 90, 98, 106 (accepting as undisputed that "a test of statistical significance . . . can result in a finding of disparate impact" for pass/fail uses); id. ¶¶ 114, 123, 134, 147 (accepting as undisputed finding of disparate impact for rank-ordering uses "assuming use of a test of statistical significance").) These admissions eliminate the existence of any factual dispute over the prima facie case.

F. The City's Arguments

1. Large Sample Sizes

Rather than attacking the accuracy of Plaintiffs' statistics, the City objects to Plaintiffs' reliance on statistical significance testing as a general matter. The City raises a number of supposed theoretical problems with such testing. The City's principal argument is that the size of the populations being tested in this case (i.e., the many thousands of applicants who took each examination) renders a statistical significance test unreliable. This is because, the City contends, the "larger the group we examine the more likely we are to find differences[]" among candidates that will cause particular individuals to fail. (Def. PF Mem. 5 (noting that "the larger the group we are examining, the more candidates who sit for the exam, the greater our likelihood that some of them will not do as well as others").)

[15] The City's Rule 56.1 Responses regarding practical significance dispute only certain assumptions of Plaintiffs' calculations, but few of these contentions are actually supported by reference to the City's own expert reports or other evidence. The disputes that are offered with such support concern the "shortfall" calculations of Dr. Siskin (see, e.g., Def. USA 56.1 ¶¶ 70-73), and the assumption of "parity" in capability and preparedness among racial and ethnic groups (see, e.g., id. ¶¶ 86-88, 95-97, 103-05, 111-13). The court addresses these assumptions below.

The City has it backwards. Rather than undermining confidence in statistical significance testing, large sample sizes make such testing more reliable. Larger sample sizes create a greater likelihood that random differences between individuals will even out among all groups, and a lower likelihood that significant differences between the performance of racial or ethnic groups will have resulted from chance. Existing precedent confirms this principle. Courts have sometimes declined to rely on statistical significance analysis when a sample size was too small. See, e.g., Lindemann 1734 ("Courts have recognized that statistical evidence often is unreliable when the sample size is small."); Pietras v. Bd. of Fire Comm'rs, 180 F.3d 468, 475 (2d Cir. 1999) (recognizing "authority holding that a disparate impact finding based solely on a sample size as small as the one presented here [i.e., 7 people] cannot stand"). Yet, the City has pointed to no cases rejecting such testing because a sample size was too large. As the Second Circuit stated in Guardians, "in cases involving large samples, 'if the difference between the expected value (from a random selection) and the observed number is greater than two or three standard deviations,' a prima facie case is established." 630 F.2d at 86 (emphasis added) (quoting Castaneda, 430 U.S. at 496 n.17).

The City's own admissions support this understanding of large sample sizes. The Federal Government has provided a helpful illustration that the City explicitly accepts as undisputed:

> Flipping a coin is a common example that illustrates why sample size should affect the number of standard deviations that is equivalent to a given disparity. Flipping a fair coin 10 times will not always result in exactly five heads and five tails; a result of six heads and four tails on ten flips would not indicate with a reasonable degree of certainty that the coin was not fair (i.e., that the disparity was not likely due to chance variation). However, if one flipped a fair coin 1,000 times, one would expect that the number of heads and tails would be close to equal, and a result of 600 heads and 400 tails would allow one to conclude with a high degree of certainty that the coin was not fair (i.e., that

> disparity between the rate at which heads came up and the rate at which tails came up was not likely do to chance variation).
>
> Put simply, with a disparity in pass rates of a given size, the bigger the sample (e.g., the more times one flips the coin, or the more applicants who take the test), the more confident one can be that the difference in pass rates in the sample is not due to chance.

(USA 56.1 ¶¶ 56-57 (internal citations omitted); see Def. USA 56.1 ¶¶ 56-57).)

Another undisputed statement, which relies on the City's own expert, further supports the reliability of statistical significance testing when large sample sizes are involved:

> With a large sample size, a test of statistical significance using 1% as the standard (i.e., concluding that there is a statistically significant disparity if there is no more than a 1% likelihood of observing a disparity so large due to chance) is better than the 80% Rule at controlling for false positives (situations in which the test used will indicate a disparity when there is no disparity) and false negatives (situations in which the test will indicate there is no disparity when there is a disparity). In other words, with a large sample size, a test of statistical significance is more likely to produce the "right" answer to the question of whether there is a non-chance disparity between the pass rates of two groups.

(USA 56.1 ¶ 64 (citing deposition of City's expert, Dr. Bobko) (emphases added); see Def. USA 56.1 ¶ 64).) These undisputed statements plainly support a conclusion that large sample sizes enhance, rather than undermine, the reliability of statistical testing. Accordingly, while the City purports to challenge the use of statistical significance testing based on sample size, the City's own admissions contradict its position.[16]

When an employment examination is used to make hiring decisions for thousands of applicants, seemingly small differences in pass rates can have a substantial effect on large groups

[16] Moreover, the Federal Government points out that Dr. Siskin has performed a recalculation of his statistical significance testing by reducing the sample sizes by 90%. (USA Mem. 18 n.13.) As the City's Bobko-Schemmer Report concedes, this recalculation based on a greatly reduced sample size does not affect the finding of statistical significance for the pass/fail uses of Exams 7029 and 2043. (See Bobko-Schemmer Report 17-18; see also Siskin Report 22, 23, 26-27, 27-28.) Although this recalculation for some of the rank-ordering disparities results in less than three units of standard deviation, the City provides no alternative statistical measure with respect to rank-ordering, and so the recalculation does not undermine the usefulness of the statistics before the court.

of people. Contrary to the City's position, therefore, it is important to rely upon statistical testing to determine whether such differences have resulted from chance or, rather, from a particular employment practice. In this case, statistical significance testing has been used to show that the disparities between groups of candidates have resulted from the challenged examinations. The City's arguments to the contrary are unavailing.

2. Perfect Parity Among Groups

The City also attacks Dr. Siskin's "shortfall" analysis, which estimates the number of minority candidates who would have passed or been appointed had the written examinations not had a discriminatory impact. The City criticizes the fact that such calculations hypothesize a world of "perfect parity" among racial or ethnic groups. (Def. PF Mem. 5.) In other words, the City argues that Plaintiffs' analyses inappropriately compare the racial disparity in test results to a hypothetical world in which racial and ethnic groups perform equally well. (See id. ("[S]tatistical significance testing will assume that all people perform at equal levels. However, we know that all individuals do not perform at the same level.").) This argument misstates the law.

First of all, the court rejects the premise that comparison to a standard of equality among groups provides an improper foundation for statistical testing under Title VII. In order to determine whether a particular employment practice has had a disparate impact on a minority group, statistical tests "ask what the results would be for the salient variable . . . if there [had been] no discrimination." Adams v. Ameritech Servs., Inc., 231 F.3d 414, 424 (7th Cir. 2000) (emphasis added). To determine what results "would be," statistical tests properly assume that racial or ethnic groups will perform equally well absent discrimination. See Smith, 196 F.3d at

366 (recognizing "null hypothesis" of no difference between compared groups).[17] Statistical significance testing relies on this assumption of equality in assessing whether disparities among groups are based upon chance, or rather, upon some other factor, such as race or national origin. See Ottaviani, 875 F.2d at 371 ("Statistical significance is a measure of the probability that a disparity is simply due to chance, rather than any other identifiable factor."); Joint Apprenticeship Comm., 186 F.3d at 117 ("a plaintiff's statistical evidence must reflect a disparity so great that it cannot be accounted for by chance"); Lindemann 94 ("Tests of statistical significance are commonly used in the social sciences to rule out chance as the cause of observed disparities."). In accordance with these principles, Plaintiffs' statistical evidence shows that the disparities in this case have not been the result of chance; instead, the disparate impact upon black and Hispanic candidates has resulted from the challenged employment practices.

The court similarly rejects the suggestion in the City's Rule 56.1 Responses that a prima facie case has not been established because disparities between white, black, and Hispanic candidates can be explained by differences in their "capability and preparedness." (See, e.g., Def. USA 56.1 ¶¶ 86-88, 95-97, 103-05, 111-13.) The City's Rule 56.1 Response suggests that the City believes black and Hispanic candidates received lower scores on its written examinations because of their lower capability and preparedness for the job of firefighter. But, if the City contends that differences in aptitude relating to the job of firefighter have led to an adverse impact on minority groups, Title VII's burden-shifting framework allows the City to

[17] By way of further example: "[i]f the relevant market is 40% African-American, for instance, one would expect 40% of hires to be African-American . . . . If the observed percentage of African-American hires is only 20%, then the statistician will compute the 'standard deviation' from the expected norm and indicate how likely it is that race played no part in the decisionmaking. Two standard deviations is normally enough to show that it is extremely unlikely (that is, there is less than a 5% probability) that the disparity is due to chance, giving rise to a reasonable inference that the hiring was not race-neutral; the more standard deviations away [from zero], the less likely the factor in question played no role in the decisionmaking process." Adams, 231 F.3d at 424.

justify the disparate impact as a matter of business necessity. At the prima facie stage, however, the question is only whether there are disparities attributable to the challenged practices, not whether the City can provide a justification for them. During this stage, the City cannot rebut the existence of disparities by claiming that they are explained by the overall "capability and preparedness" of particular groups.

Therefore, the court rejects the City's arguments about the assumptions in Plaintiffs' statistics.

3. The 80% Rule

Finally, in opposing summary judgment, the City argues that the court should rely on the 80% Rule to the exclusion of statistical significance testing. There is no support for this position. Controlling precedent holds that the 80% Rule is not an exclusive means of proof, and that alternative statistical tests should be considered. See Joint Apprenticeship Comm., 186 F.3d at 118 ("[80%] rule is not binding on courts, and is merely a 'rule of thumb' to be considered in appropriate circumstances."); see also Watson v. Fort Worth Bank & Trust, 487 U.S. 977, 995 n.3 (1988) (citing criticism of the 80% Rule, recognizing the usefulness of statistical methods in Title VII cases, and endorsing a case-by-case approach); Bew v. City of Chicago, 252 F.3d 891, 893 (7th Cir. 2001) ("The district court properly noted that the 80% guideline may be ignored when other statistical evidence indicates a disparate impact."). Moreover, the regulation containing the 80% Rule plainly endorses the use of alternative tests. It specifically states: "Smaller differences in selection rate may nevertheless constitute adverse impact, where they are significant in both statistical and practical terms. . . ." 29 C.F.R. § 1607.4D (emphasis added). Rather than supporting the use of the 80% Rule to the exclusion of statistical significance testing,

this sentence expressly contemplates alternative statistical tests as a means of showing disparate impact. This is precisely the showing Plaintiffs have made in this case.

In support of its preference for the 80% Rule, the City points to the Second Circuit's decision in Waisome. In Waisome, the plaintiffs challenged examinations used by the Port Authority in the process of promoting police officers. 948 F.2d at 1372. The standard deviation between the pass rates of white and black candidates in that case was 2.68, and the district court concluded that this was insufficient to find disparate impact because the pass rate of blacks was 87.2% the pass rate of whites. Id. at 1375. In approving this part of the district court decision, however, the Second Circuit did not find that the 80% Rule controlled the outcome of the case. Instead, the court also relied upon the fact that the calculated statistical significance, 2.68 standard deviations, was a borderline figure, and that, as a practical matter, "if two additional black candidates passed the written examination the disparity would no longer be of statistical importance." Id. at 1376.

The statistics in this case, however, are uniformly beyond 3 units of standard deviation, and, for many of the analyses performed, drastically beyond that. There are also large sample sizes, which make a finding of statistical significance more reliable. Hundreds of black and Hispanic applicants were affected by the City's written examinations. Given the practical significance of the disparities here on the actual hiring rates for black and Hispanic applicants, this case is clearly distinguishable from Waisome. Finally, it is worth noting that the Second Circuit remanded the case in Waisome because it found that there had been a showing of disparate impact. Id. at 1372.

In sum, the City has conceded the accuracy of the calculations of Plaintiffs' experts, which provide ample support for the statistical and practical significance of the disparities at issue. The City's only defense is to resort to abstract arguments relating to the nature of statistical testing in general. (See, e.g., Mar. 19, 2009 Tr. 29-30 (arguing against statistical significance testing based upon Plato's "allegory of the cave").) At the same time, the City has made admissions in its Responses to Plaintiffs' Rule 56.1 Statements that directly contradict its only purported challenges to Plaintiffs' proof. Considering its crucial admissions and the lack of legal authority for its position, the City has failed to wage a serious attack on Plaintiffs' prima facie case.

Under these circumstances, the court finds no material factual dispute relating to the prima facie case. To the extent the City purports to dispute the factual evidence presented by Plaintiffs, it has raised nothing more than metaphysical doubts about the nature of that evidence. Such doubts cannot preclude summary judgment. See Matsushita, 475 U.S. at 586. There is no dispute that Plaintiffs have satisfied a statistical standard for a prima facie case of disparate impact that has been repeatedly accepted by the Second Circuit, nor is there any dispute that they have shown that the disparity has had a substantial, practical significance for the composition of the eligibility lists and hiring of entry-level firefighters. Accordingly, the court grants summary judgment for Plaintiffs on their prima facie case of disparate impact.

IV. BUSINESS NECESSITY

While Plaintiffs have shown that the City's uses of Written Examinations 7029 and 2043 resulted in a disparate impact upon black and Hispanic candidates, the City may defend against Title VII liability by showing that those uses were justified by legitimate business and job-related

considerations.[18] The City bears the burden of making this showing.[19] See Gulino v. New York State Educ. Dep't, 460 F.3d 361, 385 (2d Cir. 2006). In Gulino, the Second Circuit explained the business necessity defense as follows:

> [T]he basic rule has always been that "discriminatory tests are impermissible unless shown, by professionally acceptable methods, to be predictive of or significantly correlated with important elements of work behavior which comprise or are relevant to the job or jobs for which candidates are being evaluated." This rule operates as both a limitation and a license for employers: employers have been given explicit permission to use job related tests that have a disparate impact, but those tests must be "demonstrably a reasonable measure of job performance."

460 F.3d at 383 (quoting Albemarle Paper Co. v. Moody, 422 U.S. 405, 431, 426 (1975)).

In setting forth its analysis, this court first reviews the process by which the City created the two challenged examinations. The court then addresses a motion to strike post-discovery submissions made by the City relating to its business necessity defense. Finally, the court sets

[18] "Though the terms 'business necessity' and 'job related' appear to have semantic differences, they have been used interchangeably by the courts." Gulino v. New York State Educ. Dep't, 460 F.3d 361, 382 (2d Cir. 2006).

[19] The basic disparate impact burden-shifting framework suggests that the determination of whether there is a sufficient disparity is limited to the prima facie case, and that, once that burden is satisfied, the City must show job-relatedness. Nevertheless, Gulino and Robinson unmistakably state that, following a prima facie showing, a defendant may still "directly attack plaintiff's statistical proof by pointing out deficiencies in data or fallacies in the analysis." Gulino, 460 F.3d at 382; see also Robinson, 267 F.3d at 161. At that stage, "[t]o successfully contest the plaintiffs' statistical evidence . . the employer has to convince the fact finder that its numerical picture is more accurate, valid, or reliable than the plaintiffs' evidence." Robinson, 267 F.3d at 161 (internal quotation marks and alteration omitted). The employer bears the burden of this showing.

The parties do not explicitly address this component of the City's burden, as their arguments relating to statistics are made entirely in the context of the prima facie case. However, the City does argue that "trial courts have discretion to decide whether to use the 80% Rule of statistical significance testing in deciding Title VII claims," and that the court "should choose [to] employ the 80% Rule . . . in the instant matter." (Def. PF Mem. 3, 4.) Thus, the City asks the court to proceed to trial on the issue of whether "its numerical picture is more accurate, valid, or reliable than the plaintiffs' evidence." Robinson, 267 F.3d at 161. But, as discussed above, the court need not conduct a trial to resolve the question of whether the 80% Rule should be used to the exclusion of Plaintiffs' statistical tests, when the controlling legal authority has already answered that question. Moreover, to the extent that the City's position attempts to show "fallacies in the analysis" of Plaintiffs' experts, see Gulino, 460 F.3d at 382, its arguments have come entirely in the form of unsupported attacks on the usefulness of statistical significance testing in general. Besides these attacks, discussed and rejected above, the City has completely conceded Plaintiffs' statistical picture. With no factual dispute on which to conduct a trial, summary judgment is appropriate.

out and applies the relevant Second Circuit standard for assessing the challenged examinations. The court concludes that the City has not met that standard.

A. Creation of Challenged Examinations

The City used the same development process for Written Exams 7029 and 2043. For Written Exam 7029, a "Test Development Report" was prepared by Matthew Morrongiello ("Morrongiello"), a Tests and Measurement Specialist in the City's DCAS. (See Levy Decl. Ex. EE ("Test Development Report").) Alberto Johnston ("Johnston") of DCAS was "primarily responsible" for developing Exam 2043, and testified that he "was told that we probably can use the old job analysis . . . from [Exam] 7029 . . . ." (Fraenkel Decl. Ex. 6. ("Johnston Dep."), at 17-18, 19.) Accordingly, the same Test Development Report was relied upon in the City's development of Written Exam 2043. (See Int. 56.1 ¶ 12; see also id. ¶¶ 85-86; Levy Decl. Ex. CC, at 5 (Admission # 11); Fraenkel Decl. Ex. 10 ("Patitucci Dep."), at 208-09.)[20]

The Test Development Report is a twelve-page document with nine appendices that sets out the process by which the City arrived at the abilities it intended to evaluate on the written examinations, as well as the number of questions it would devote to each ability. The salient features of the process, as set out in that Report, are not in dispute.[21] The Test Development

[20] The City also points to a draft report by Dr. Frank Landy entitled, "Job Analysis and Written Examination Development For the City of New York Firefighter Examination No. 0084" dated December 18, 1992. (See Frankel Decl. Ex. 5 ("Landy Report"); see also Def. BN Mem. 5 (relying on the Landy Report).) The Landy Report was prepared in the course of developing an earlier examination, Written Examination 0084. (See Test Development Report 4.) Morrongiello, who was responsible for developing Exam 7029 and the Test Development Report, stated in his deposition that he relied on the Landy Report "to a degree." (See Fraenkel Decl. Ex. 22 ("Morrongiello Dep."), at 478-79.) Specifically, he stated that he used the task list in that report as a starting point from which to begin developing Exam 7029. (See id. at 478.) As indicated below, the Test Development Report supports this assertion.

[21] The Intervenors' Rule 56.1 Statement sets out the various steps in the process described in the Test Development Report, relying largely on the Report but supplementing its statements with other support in the record. (See, e.g., Int. 56.1 ¶¶ 87-90.) In response, the City has denied most of the assertions and cited the Test Development Report

Report states that a series of meetings were held, pursuant to which DCAS concluded that it would conduct a new, "comprehensive" job analysis, determine how many people would be needed for a "job analysis survey," and convene a panel of 8 to 12 incumbent firefighters to review the task and ability lists and the job analysis questionnaire.[22] (See Test Development Report 4.) The basic plan for the process was as follows: first, the tasks and abilities most relevant to the job of firefighter would be determined; second, the relative importance of these tasks and abilities would be assessed; third, clusters of tasks would be matched up with the abilities needed to perform them; and finally, a test would be created to evaluate the identified abilities in the proper proportions. The court briefly reviews this process, as set forth in the Test Development Report.

1. Deriving a List of Tasks and Abilities

In order to familiarize himself with the job of firefighter, Morrongiello conducted interviews with six incumbent firefighters. (Id. at 5.) He used the results of these interviews, coupled with the task list used to develop a prior examination, Exam 0084, to create an "updated task list" reflecting the tasks performed by firefighters. (Id.) Morrongiello then compiled a list of 21 cognitive abilities, derived from "Fleishman's ability list." (See Test Development Report 5; Levy Decl. Ex. X, at 129.) Morrongiello convened a focus group of ten firefighters who reviewed the task and ability lists, and, based on the focus group, "suggested changes to the

---

"for a full and accurate statement of its contents and the steps taken by [] Morrongiello." (See, e.g., Def. Int. 56.1 ¶¶ 87-90.) There is no dispute about the process, but only about the significance of the steps taken.

[22] The written examinations were not intended to include any measure of physical ability, because these would be tested on a separate physical examination, presumably the PPT. (Test Development Report 4.) The Test Development Report states an intention to include questions on the job analysis questionnaire about Oral Comprehension, Written Comprehension, Oral Expression, and Written Expression. (Id.)

proposed task list so it more accurately reflected the current job of Firefighter." (Test Development Report 5.)

The focus group also reviewed a Job Analysis Questionnaire ("JAQ") that was distributed to 195 incumbent firefighters. (See id. at 5, 6, 10.) The JAQ asked the firefighters to assess whether 196 listed tasks (grouped into 21 specific task clusters, plus a miscellaneous cluster) were "4. Critical," "3. Important," "2. Somewhat important," or "1. Not [] important" to the job of firefighter, and, similarly, whether 21 listed abilities were "4. Critical," "3. Important," "2. Somewhat important," or "1. Not relevant" to the firefighter job. (Id. at 10 & app. E (JAQ test results).) The responses to the JAQ by 192 of the 195 surveyed firefighters, excluding 3 defective responses, were used to hone the list of abilities down to 18, and the list of tasks down to 111. (Id. at 10.) This was done by removing those tasks and abilities which did not receive at least an average score of 2.5, corresponding to a rating between "important" and "somewhat important." (Id.)

2. Linking Tasks With Abilities

With these results in hand, Morrongiello assembled twelve firefighters into a "Linking Panel," whose purpose was to "rate—or link—the task clusters to the abilities." (Id.) The specific "clusters" of tasks important to the job of entry-level firefighter were:

- Initial Response to Incidents/Driving (tasks that "occur between receiving an alarm and initial fire fighting or emergency activities, including driving apparatus to and from various points");
- Size Up (tasks that involve "evaluating the fire or incident scene to determine actions which should initially be taken and obtaining information needed for evaluation");
- Ladder Operation (tasks that involve "stabilizing ladder trucks and elevating and operating aerial ladders and platforms in order to rescue victims, provide access for ventilation, operate master stream devices, etc.");

- Climbing and Portable Ladder Activities (tasks that involve "climbing ladders, stairs and fire escapes, and raising and setting up portable ladders");
- Building Entry (tasks that involve "prying open or breaking through doors or otherwise entering buildings in order to search for and rescue victims and provide access to the fire for offensive fire fighting, using axes, halligan tool, hooks, rabbit tools, sledge hammers, power saws, and other tools");
- Search (tasks that involve "searching fire or assigned area in order to locate victims and to obtain further information about fire, following standard search procedures");
- Rescue (tasks that involve "assisting, carrying or dragging victims from emergency area by means of interior access (stairs, hallways, etc.) or, if necessary, by ladders, fire escapes, platforms, or other means of escape");
- Ventilation (tasks that involve "opening or breaking open windows, chopping or cutting holes in roofs, breaking through walls or doors, and hanging fans in windows or doors to remove heat, smoke and gas from burning buildings");
- Supplies Water for Hose Operation (tasks that involve "connecting or hooking up engine to fire hydrant and operating pumps to supply water in appropriate pressure and volume for fire fighting, using hydrant wrenches, couplings, hoses, spanner wrenches, and other tools");
- Hose Operations During Extinguishment (tasks that involve "stretching line to fire scenes and delivering water to scene of fire");
- Overhaul (tasks that involve "opening up walls and ceilings, cutting or pulling up floors and moving or turning over debris, in order to check for hidden fires which could rekindle or spread, using hooks, axes, saws and pitchforks");
- Salvage (tasks that involve "moving and covering furniture, appliances, merchandise and other property, and covering holes in buildings and redirecting or cleaning up water in order to minimize damage, using plastic and canvas covers, ropes, staple guns, mops, squeegees, and other tools");
- Clean Up/Pick Up (tasks that involve "picking up and returning equipment to vehicle and rolling up or folding up hose, so that the company can go back in service");
- Equipment Maintenance (tasks that involve "inspecting, cleaning, and maintaining apparatus, equipment carried on the apparatus, and personal gear and equipment");
- Inspection of Buildings/Hydrants (tasks that involve "inspecting buildings for code violations or hazards on a periodic basis or during the course of activities, and inspecting hydrants for operational use");
- Extrication (tasks that involve "extricating victims from vehicles, cave-ins, collapsed buildings or other entrapments in order to save lives, using shovels, torches, drills, pry bars, saws, jacks, hurst tools, air bags, and other equipment");

- Providing Medical Assistance (tasks that involve "providing first aid and direct medical assistance to persons requiring emergency attention");
- Elevator Related Tasks (tasks that involve "controlling elevators and rescuing persons from stalled elevator cars");
- Training (tasks that involve "participating in drills which simulate important fire or rescue activities, and attending lectures or formal training");
- Watch Duties (tasks that involve "standing watch to receive incoming alarms and information, answering phones, and monitoring access to the station house");
- Station Duties and Chores (tasks that involve "performing routine housekeeping chores or 'committee work'"); and
- Miscellaneous (tasks that involve "miscellaneous tasks").

(See id. app. E, at USA000371 – USA000380; see also Levy Decl. Ex. FF ("Linking Panel Worksheet").) The 18 abilities to which members of the Linking Panel were supposed to "link" these tasks were:

- Oral Comprehension (the ability to "understand spoken English words and sentences");
- Written Comprehension (the ability to "understand written sentences and paragraphs");
- Oral Expression (the ability to "use English words or sentences in speaking so that others will understand");
- Written Expression (the ability to "use English words or sentences in writing so that others will understand");
- Fluency of Ideas (the ability to "produce a number of ideas about a given topic");
- Originality (the ability to "produce unusual or clever ideas about a given topic or situation," and to "invent creative solutions to problems or to develop new procedures for situations in which standard operating procedures do not apply");
- Memorization (the ability to "remember information, such as words, numbers, pictures and procedures. Pieces of information can be remembered by themselves or with other pieces of information");
- Problem Sensitivity (the ability to "tell when something is wrong or is likely to go wrong. It includes being able to identify the whole problem as well as elements of the problem");
- Deductive Reasoning (the ability to "apply general rules to specific problems to come up with logical answers. It involves deciding if an answer makes sense");

- Inductive Reasoning (the ability to "combine separate pieces of information, or specific answers to problems, to form general rules or conclusions. It involves the ability to think of possible reasons for why things go together");
- Information Ordering (the ability to "follow correctly a rule or set of rules or actions in a certain order. The rule or set of rules used must be given. The things or actions to be put in order can include numbers, letters, words, pictures, procedures, sentences, and mathematical or logical operations");
- Speed of Closure ("involves the degree to which different pieces of information can be combined and organized into one meaningful pattern quickly. It is not known beforehand what the pattern will be. The material may be visual or auditory");
- Flexibility of Closure (the ability to "identify or detect a known pattern (like a figure, word, or object) that is hidden in other material. The task is to pick out the disguised pattern from the background material");
- Spatial Orientation (the ability to "tell where you are in relation to the location of some object or to tell where the object is in relation to you");
- Visualization (the ability to "imagine how something would look when it is moved around or when its parts are moved or rearranged. It requires the forming of mental images of how patterns or objects would look after certain changes, such as unfolding or rotation. One has to predict how an object, set of objects, or pattern will appear after the changes have been carried out");
- Perceptual Speed ("involves the degree to which one can compare letter, numbers, objects, pictures, or patterns, quickly and accurately. The things to be compared may be presented at the same time or one after the other. This ability also includes comparing a presented object with a remembered object");
- Selective Attention (the ability to "concentrate on a task one is doing. This ability involves concentrating while performing a boring task and not being distracted"); and
- Time Sharing (the ability to "shift back and forth between two or more sources of information").

(See Test Development Report app. F, at USA000397 – USA000398; id. app E, at USA000381 – USA000382; id. app. F, at USA000384 – USA000394.)[23]

[23] The Test Development Report, included at Exhibit EE of the Levy Declaration and at Exhibit 4 of the Fraenkel Declaration, appears to have certain pages missing from the appendices. Moreover, the Final Task List sets out 14 of the abilities, although a page seems to be missing from this part of the appendix. The Linking Panel Checklist, however, includes ratings for 18 of the abilities, and the JAQ questionnaire includes descriptions of these 18 abilities. The description of 18 abilities is derived from Appendices E and F of the Report.

The goal of the Linking Panel was to match up these 18 abilities with the 21 identified task clusters. (See Test Development Report app. F ("You will . . . be asked to rate the importance of each of the eighteen abilities for the performance of each of the twenty-one task clusters."); see also Linking Panel Worksheet.) Members of the Linking Panel each had to come up with a rating to reflect how important each ability was to each cluster. (See id.) This rating was either "Critical to the performance of the task cluster," "Important to the performance of the task cluster," "Somewhat important to the performance of the task cluster," or "Not relevant to the performance of the task cluster." (See id.)

Although this appears to have been the only step in the process intended to capture the relationship between the tasks of a firefighter and the abilities tested on the written examination, the Test Development Report does not explain how or why particular tasks were matched with particular abilities. In fact, before performing this task, Linking Panel members were not given any explanation about the meaning of the ratings they were supposed to provide. (Int. 56.1 ¶ 95.) No statistical analyses were conducted to confirm the reliability of the ratings or the agreement in ratings among panel members. (See Morrongiello Dep. 299-300.)

Although the Linking Panel matched the 21 clusters to 18 abilities, only nine of the 18 abilities were deemed "testable" in a written multiple-choice format: Written Comprehension, Written Expression, Memorization, Problem Sensitivity, Deductive Reasoning, Inductive Reasoning, Information Ordering, Spatial Orientation, and Visualization. (Test Development Report 11.) The two abilities which incumbent firefighters rated highest in importance—Oral Comprehension and Oral Expression—were not among those tested, because "structured interviews" with thousands of candidates (which would help evaluate oral abilities) would not

have been feasible. (See, e.g., Fraenkel Decl. Ex. 11 ("Patitucci II Dep."), at 131-32, 274-75; see also Int. 56.1 ¶ 106; Def. Int. 56.1 ¶ 106.) Regarding the other seven omitted abilities, Morrongiello stated that he "didn't do anything specific to determine" whether or not they were testable, and that he was "going by . . . standard operating procedure at that time in our unit . . . that these abilities" would not be tested. (Morrongiello Dep. 443.)

To determine how many examination questions would be devoted to each ability, the "average importance ratings for each of the nine testable abilities within each cluster were determined from the individual ratings given by the linking panel." (Test Development Report 11; see also id. app. G (setting out column with average importance rating of ability "to task cluster").) The panel multiplied the average of each importance rating by the rating that the JAQ questionnaires had given to the ability. (Id. at 11.) An average rating was then calculated for each ability—that rating was "pro-rated" and rounded based on an 85-question, multiple-choice test. (Id.)

The result of this process was a test intended to evaluate nine abilities as follows: Written Comprehension (9 questions), Written Expression (6 questions), Memorization (11 questions), Problem Sensitivity (12 questions), Deductive Reasoning (9 questions), Inductive Reasoning (9 questions), Information Ordering (11 questions), Spatial Orientation (10 questions), and Visualization (8 questions). (See Test Development Report, at USA000404; Levy Decl. Ex. M, at 4 (Admission # 30).) The parties agree that all of these are "cognitive" abilities. (See Levy Decl. Ex. M, at 4, 6 (Admission ## 29, 36).)

3. Test Construction

The next step in the process was to construct a written examination based upon the job analysis. For Exam 7029, one Lieutenant and four firefighters were "given training in how to write exams," were placed on a panel, and then wrote the examination. (Test Development Report 11.) A "Review Panel" was assembled to review the questions on the examination (Int. 56.1 ¶ 124; Test Development Report 12), although it is unclear what this panel reviewed the questions for. One thing that the reviewers did not consider was whether "each [question] measured the ability it was originally designed to measure." (Int. 56.1 ¶ 124.) No analysis of the reading level of the examination was conducted. (Int. 56.1 ¶ 128.)

This test-writing process was essentially the same for Exam 2043. (See Johnston Dep. 23-28.)[24]

B. Motion to Strike

The determination of whether an employment test is job-related relies heavily upon expert testimony. Although discovery in this case closed in October 2008, the City submitted two new declarations containing expert assertions with its summary judgment papers in February 2009: a declaration from the City's expert, Dr. Schemmer (Fraenkel Decl. Ex. 2 ("Schemmer Decl.") and a declaration from Dr. Catherine Cline, who participated in the development of Exam 6019, administered after Exams 7029 and 2043 (Fraenkel Decl. Ex. 3 ("Cline Decl.")). Plaintiffs have moved to strike these declarations. (See Docket Entries ## 273, 274.) For the reasons that follow, the court grants the motion in part and denies it in part.

[24] While Exam 7029 was developed by Morrongiello, Exam 2043 was developed by Johnston. The same job analysis and test plan were used. (See Int. 56.1 ¶ 12.) According to Johnston, a similar process was used to draft the questions, including the convening of a panel of incumbent firefighters to write the test questions for Exam 2043. (See Johnston Dep. 23-28, 122-24.)

The Intervenors argue that the court should strike these declarations because the deadline for submitting expert reports was January 21, 2008, and all expert and fact discovery closed on October 31, 2008. (See Declaration of Richard Levy dated March 4, 2009 (Docket Entry # 274).) The Intervenors point out that Federal Rule of Civil Procedure 26(a) requires a written report with a complete statement of an expert witness' opinions, including the reasons for them, and that Rule 26(e) requires supplementation of that report "in a timely manner." (See Memorandum of Law In Support of Motion to Strike (Docket Entry # 274) ("Strike Mem.") 2-3.) They further argue that Rule 37(c)(1) prevents a party from relying on information it did not disclose in accordance with Rules 26(a) and (e). (See id. at 3.) Because the declarations offer new expert opinions in violation of the discovery rules, the Intervenors ask the court to strike them.

Under Rule 26(a)(2)(B) of the Federal Rules of Civil Procedure, expert testimony must be accompanied by a written report which shall contain, inter alia, "a complete statement of all opinions the witness will express and the basis and reasons for them," "the data or other information considered by the witness in forming them," and "any exhibits that will be used to summarize or support them." A party must make these disclosures "at the times and in the sequence that the court orders." Fed. R. Civ. P. 26(a)(2)(C). Rule 37(c)(1) states that if a party fails to abide by these requirements, "the party is not allowed to use that information . . . to supply evidence on a motion, at a hearing, or at a trial, unless the failure was substantially justified or is harmless." The Second Circuit has construed the language in Rule 37(c)(1) to provide discretion to preclude evidence if "the trial court finds that there is no substantial

justification and the failure to disclose is not harmless." Design Strategy, Inc. v. Davis, 469 F.3d 284, 294 (2d Cir. 2006).

The parties have had ample time to conduct expert discovery. As Magistrate Judge Roanne Mann's Scheduling Orders make clear, the City was required to make expert disclosures on business necessity by January 7, 2008. (See Scheduling Order (Docket Entry # 30) ¶ 4.) That deadline was extended to January 21, 2008. (See Revised Schedule (Docket Entry # 66) ¶ 4.) The schedule for the City's expert depositions was amended several times, and the deadline was last scheduled for the end of March 2008. (See Scheduling Order (Docket Entry # 30) ¶ 5; Revised Schedule (Docket Entry # 66) ¶ 5; Modified Scheduling Order (Docket Entry # 86) ¶ 5.) A schedule for Plaintiffs' rebuttal on business necessity was also ordered by Judge Mann, with all expert and fact discovery to conclude on October 31, 2008. (See Modified Scheduling Order (Docket Entry # 181).) The October deadline was set at the direction of this court to ensure that all discovery would be completed by then. (See April 10, 2008 Tr. 15-16.)

Clearly, submitting new expert opinions after the close of discovery violates the discovery rules. The City does not dispute this principle, instead arguing that it is simply not making new expert disclosures. Regarding the Cline Declaration, the City states that Dr. Cline is not being offered as an expert witness, but is, rather, only being offered as a fact witness to correct certain remarks about Exam 6019 that were made in Intervenors' summary judgment papers. (See Memorandum of Law in Opposition to Motion to Strike (Docket Entry # 276) ("Strike Opp.") 4, 6 (contending that Dr. Cline is "not being offered as an expert," but simply as "a fact witness concerning her work" developing Exam 6019).) Regarding the Schemmer

Declaration, the City states that its submission merely provides further "detail" on issues already opined on by Dr. Schemmer. (Id. at 2-3.)

The City's position on both counts is disingenuous. First, the Cline Declaration consists primarily of expert opinions about the validity of Exams 7029 and 2043. Except in a few places, these assertions are based upon specialized knowledge of the art of test construction and validation, rather than the personal knowledge of a lay witness. Although it appears that Dr. Cline might have qualified to serve as an expert—had the City offered her, following the proper procedures—the City has chosen not to do so. If Dr. Cline is not offered as an expert, she may not opine about the validity of Exams 2043 and 7029, about which she has no personal knowledge. See Fed. R. Evid. 701; United States v. Rigas, 490 F.3d 208, 224 (2d Cir. 2007) ("Rule 701(c), which prohibits testimony from a lay witness that is 'based on scientific, technical, or other specialized knowledge,' is intended 'to eliminate the risk that the reliability requirements set forth in Rule 702 will be evaded through the simple expedient of proffering an expert in lay witness clothing.'").

Based on the court's review of the Cline Declaration, it is clear that it is nothing more than an expert declaration submitted following the close of expert discovery. It would be prejudicial to Plaintiffs to have to address these new expert opinions from Dr. Cline, who was never offered as an expert. The Cline Declaration shall therefore be stricken. Nevertheless, those portions of the Declaration that simply clarify, as a factual matter within Dr. Cline's personal experience, the preparation of Exam 6019, will not be stricken. (See id. ¶¶ 11 (first six sentences based on personal knowledge), 12 (first sentence based on personal knowledge), 16 (first sentence based on personal knowledge).) Exam 6019 is not at issue in this litigation, and it

would be harmless to supplement the record regarding that examination. The court need not strike those portions of the Cline Declaration.

Second, the court rejects the City's argument that the Schemmer Declaration offers no new expert opinions. That Declaration contains numerous paragraphs directly addressing issues for which the City has offered no reference to a timely report or disclosure. (See, e.g., Schemmer Decl. ¶¶ 3-14.) The numerous conclusory assertions in the Schemmer Declaration suggest that they were constructed to fill holes in the evidence that the City failed to gather during discovery, and to rebut analyses presented over a year ago in Plaintiffs' expert reports. See Point Prods. A.G. v. Sony Music Entertainment, Inc., No. 93-cv-4001(NRB), 2004 WL 345551, at *9 (S.D.N.Y. Feb. 23, 2004) ("To accept the contention that the new affidavits merely support an initial position when they in fact expound a wholly new and complex approach designed to fill a significant and logical gap in the first report would eviscerate the purpose of the expert disclosure rules."). Dr. Schemmer's largely conclusory and unsupported statements strongly suggest an attempt by the City to "sandbag" its opponents with new opinions designed to defeat summary judgment. See Disability Advocates, Inc. v. Paterson, No. 03-CV-3209 (NGG)(MDG), 2008 WL 5378365, at *11 (E.D.N.Y. Dec. 22, 2008) ("The purpose of [the disclosure rules] is to prevent the practice of 'sandbagging' an opposing party with new evidence."). It would be prejudicial to Plaintiffs to have to address these assertions at this point in the litigation.

The court will not consider new, conclusory opinions by Dr. Schemmer. The City was aware of its burden to demonstrate business necessity during the discovery process, and it is bound by the analysis and opinions offered by Dr. Schemmer during that time. See Wechsler v.

Hunt Health Sys., Ltd., 381 F. Supp. 2d 135, 156 (S.D.N.Y. 2003). Indeed, the City does not even attempt to argue that new evidence should be considered, and simply pretends that the Schemmer Declaration offers no new opinions. To allow such new evidence to be presented would undermine the purpose of the discovery rules, circumvent the discovery schedule that was ordered by the court, and prejudice Plaintiffs. Accordingly, the court strikes the Schemmer Declaration in its entirety.

The court now turns to the merits of the City's business necessity defense.

C. Guardians and the Validity of Employment Tests

To be considered job-related, an employment examination must be properly "validated," and the Second Circuit has identified two sources that help determine their validity: (1) "the testimony of experts in the field of test validation" and (2) "the Equal Employment Opportunity Commission's 'Uniform Guidelines on Employee Selection Procedures' ("EEOC Guidelines")." Gulino, 460 F.3d at 382 (citing 29 C.F.R. §§ 1607.1-1607.18). Each source is important to a court's decision. As the Second Circuit stated in Gulino, while courts "must take into account the expertise of test validation professionals," they "must also remain aware that reliance upon the findings of experts in the field of testing should be tempered by the scrutiny of reason and the guidance of Congressional intent." Id. (internal citation and quotation marks omitted).[25] And, although courts must "approach the [EEOC] Guidelines with the appropriate mixture of deference and wariness, thirty-five years of using these Guidelines makes them the primary

[25] The Intervenors' 56.1 Statement describes in detail many of the findings and opinions of Plaintiffs' experts, Drs. David P. Jones and Laeetta M. Hough, Dr. Joel P. Wiesen, and Dr. Irwin L. Goldstein, as well as the opinions of City experts, Drs. Philip Bobko and F. Mark Schemmer. Accompanying this submission are excerpts from the expert reports of Drs. Jones and Hough (Levy Decl. Ex. U ("Jones-Hough Report")), Dr. Siskin (Levy Decl. Ex. V ("Siskin II Report")), Dr. Wiesen (id. Exs. R, Z ("Wiesen Report")), Dr. Goldstein (id. Ex. DD ("Goldstein Report")), and Drs. Bobko and Schemmer (Fraenkel Decl. Ex. 1 ("Bobko-Schemmer Report")), as well as deposition testimony.

yardstick by which we measure defendants' attempt to validate" employment tests. Id. at 384 (internal citation and quotation marks omitted).

The governing case in this Circuit for assessing the validity of employment tests is Guardians Association of the New York City Police Department, Inc. v. Civil Service Commission, 630 F.2d 79, 82 (2d Cir. 1980). See Gulino, 460 F.3d at 385 ("Guardians is still the law in this Circuit."). Guardians involved a test administered by the City in 1979 to over 36,000 applicants for positions in the New York City Police Department. "The exam was developed by a fairly elaborate two-stage process," with stage one involving a job analysis with input from, among others, numerous panels of police officers and questionnaires to thousands of police officers, and stage two involving additional panels and test-question revision from police experts and the New York City Department of Personnel. 630 F.2d at 83-84. The examination had a disparate impact, and the principal issue on appeal was "whether the defendants have rebutted the plaintiffs' prima facie case by showing that its test was job-related." Id. at 88.

In assessing whether the employment test was job-related, Guardians recognized that the EEOC Guidelines set forth a "sharp distinction" between "tests that measure 'content'—i.e., the 'knowledges, skills or abilities' required by a job—and tests that purport to measure 'constructs'—i.e., the 'inferences about mental processes or traits, such as 'intelligence, aptitude, personality, commonsense, judgment, leadership and spatial ability.'" Gulino, 460 F.3d at 384 (quoting Guardians, 630 F.2d at 91-92) (emphases added). "To demonstrate 'content validity,' the employer must introduce data 'showing that the content of the selection procedure is representative of important aspects of performance on the job for which the candidates are to be evaluated.'" Id. at 384 n.23 (quoting 29 C.F.R. § 1607.5B and citing 29 C.F.R. § 1607.14C).

"To demonstrate 'construct validity' on the other hand, the employer must introduce data 'showing that the procedure measures the degree to which candidates have identifiable characteristics which have been determined to be important in successful performance in the job for which the candidates are to be evaluated.'" Id. (quoting 29 C.F.R. § 1607.5B and citing 29 C.F.R. § 1607.14D).

Guardians criticized the sharp distinction between "content validity" and "construct validity." The court observed that, under the EEOC Guidelines, "content validation is generally much easier to achieve than construct validation," even though the test types differ more in degree than in kind. Id. at 384. This is because "content and construct represent a continuum that 'starts with precise capacities and extends to increasingly abstract ones.'" Id. (quoting Guardians, 630 F.2d at 93). Because of the difficulty in showing "construct validity," the court observed that "a conclusion that construct validation is required would often decide a case against a test-maker, once a disparate racial impact has been demonstrated." Guardians, 630 F.2d at 92.

In response to these concerns, the court tempered the standards set forth in the EEOC Guidelines with a functional approach to test validation. See Gulino, 460 F.3d at 386 n.25. Guardians established a five-part test to determine the content validity of an employment test, which is flexible enough to encompass concepts of construct validity:

(1) the test-makers must have conducted a suitable job analysis;
(2) they must have used reasonable competence in constructing the test itself;
(3) the content of the test must be related to the content of the job;
(4) the content of the test must be representative of the content of the job; and
(5) there must be a scoring system that usefully selects from among the applicants those who can better perform the job.

Id. at 384-85 (quoting Guardians, 630 F.2d at 95).

The first two requirements relate to the "quality of the test's development," while the final three "are more in the nature of standards that the test, as produced and used, must be shown to have met." Guardians, 630 F.2d at 95. Gulino instructs that the Guardians approach "has the advantage of tracking the [EEOC] Guidelines standards while still allowing the courts to take a more functional approach to the analysis," and frees "the courts from having to draw sharp distinctions between 'content' and 'construct' or 'knowledge' and 'ability.'" 460 F.3d at 385 (quoting Guardians, 630 F.2d at 93-94).

The parties do not dispute that Guardians provides the appropriate standard by which to evaluate Written Exams 7029 and 2043.

D. Application of Guardians

The court must determine whether the City has offered sufficient evidence to create a disputed issue of material fact that Written Examinations 7029 and 2043 are job-related under Guardians. The basic question before the court is whether the examinations selected candidates who would be better firefighters. See Guardians, 630 F.2d at 88. The more specific question is whether the City has met the detailed requirements of test validation set out by Guardians. See Lindemann 151 ("Guardians contains an unusually complete discussion of the details of test validation . . . [and] the validation criteria set forth in Guardians are ones that employers should attempt to satisfy in comparable situations.").

The court rejects the City's assertion that there are material factual disputes sufficient to preclude summary judgment on job-relatedness. Even considered in the light most favorable to the City, the undisputed evidence paints an extremely troubling picture of the test construction

process and the content that the City sought to test. Even under the summary judgment standard, the City has failed to meet its burden to show that its reliance on the challenged examinations was warranted by a valid business justification. For each of the Guardians requirements, the City's arguments are riddled with serious defects, and the facts it presents patently fail to satisfy the demands of test validation. Insufficient evidence is presented for a reasonable fact finder to conclude that the challenged examinations were related to the job of a firefighter and relied upon as a matter of business necessity. The court finds itself compelled to grant summary judgment for Plaintiffs.

As the court sets forth below, the City's evidence fails to show valid test construction under Guardians' first and second requirements, and fails to establish appropriate test content under Guardians' fourth requirement. The City's showing on Guardians' third requirement demonstrates only a minimal relationship between the content of its examinations and the content of the job of firefighter. These serious failings culminated in the City's decision to use the problem-riddled examinations to impermissibly fail and arbitrarily rank firefighter candidates. The imposition of these scoring devices, based upon the results of poorly constructed examinations, means that the City has failed to meet the fifth Guardians requirement. The examinations were simply unable to "select from among the applicants those who can better perform the job." Guardians, 630 F.2d at 95. The recurrence of severe deficiencies at every step of the court's review destroys any pretense that the challenged examinations had "a manifest relationship to the employment in question." Albemarle Paper Co., 422 U.S. at 425 (citation omitted).

1. Job Analysis

"According to the [EEOC] Guidelines, a job analysis involves an assessment 'of the important work behavior(s) required for successful performance and their relative importance.'" Guardians, 630 F.2d at 95 (quoting 29 C.F.R. § 1607.14C(2)). In Guardians, the court concluded that the City's extensive job analysis adequately identified 42 important work tasks or behaviors. Specifically, the "work behaviors involved in being a police officer were identified by extensive interviewing, and subjected to serious review . . . ." Id. at 95; see also id. at 83 (describing process by which New York City Personnel Department identified tasks performed by police officers and a panel of police officers honed the tasks identified). The relative importance of the 42 tasks were assessed by "means of an extensively distributed questionnaire" used to rank them. Id; see also id. at 83 (noting that over 2,600 police officers answered questionnaires to help rank the tasks' importance).

Nevertheless, Guardians deemed the overall job analysis to have been of "questionable sufficiency." Id. at 96. In transforming the 42 tasks into the corresponding "knowledge, skills or abilities necessary to the effective performance" of those tasks, "no effort was made to explain the relationship between any of the . . . abilities and the 42 job tasks from which they were ostensibly derived." Id. Because of this shortcoming, the Second Circuit cautioned that, "[o]nly if the relationship of abilities to tasks is clearly set forth can there be confidence that the pertinent abilities have been selected for measurement." Id. at 98.

This deficiency in the City's job analysis is also present here. As in Guardians, the City conducted a job analysis aimed at identifying the tasks of an entry-level firefighter. The City took measures to develop an extensive task list based on panels and job questionnaires with

incumbent firefighters. (See Test Development Report 4-10.) It then used those results to pare down the list to 21 specific task clusters and 18 necessary abilities. (See id. app. E, at USA000371 – USA000393; id. app. E, at USA000381 – USA000382; id. app. F, at USA000397 – USA000398; see also Linking Panel Worksheet.)

However, as in Guardians, the City has offered no evidence of "the relationship of abilities to tasks." 630 F.2d at 96. The absence of such evidence undermines the court's confidence "that the pertinent abilities have been selected for measurement." Id. Looking at the 21 task clusters—including such categories as Ladder Operation, Climbing and Portable Ladder Activities, Building Entry, Search, Rescue, Ventilation, Hose Operations During Extinguishment, and Extrication—it is not apparent how they relate to the nine specific abilities identified by the City for testing. Indeed, the City has not offered any explanation or documentation indicating how the task clusters relate to the nine abilities. Cf. M.O.C.H.A. Soc'y, Inc. v. City of Buffalo, No. 98-CV-99C(JTC), 2009 WL 604898, at *14 (W.D.N.Y. Mar. 9, 2009) (noting that procedures, including linking tasks and abilities, were "painstakingly documented").

Not only is there an absence of evidence supporting the relationship between tasks and abilities, but there is also strong evidence that no such relationship exists. Deposition testimony of Linking Panel members shows a considerable degree of confusion about the process and about the definitions of abilities the members were supposed to evaluate. For example, one member testified that he "probably didn't know" what Inductive Reasoning meant when providing a rating, "so I gave it a two since I didn't know what it was, quite frankly." (Frankel Decl. Ex. 20, at 66-67.) When presented with a definition, he was able to explain its importance, but he also

stated that he had not been given definitions of the abilities at the time he made his linking determinations. (Id.) This panel member also stated that "I don't believe I really knew what deductive reasoning was at the time of the examination." (Id. at 72.) Other members testified to being unsure of the meanings of various abilities. (See Fraenkel Decl. Ex. 21, at 57 (Problem Sensitivity)), id. at 57-58 (Deductive Reasoning), id. at 58 (Inductive Reasoning), id. at 58-59 (Information Ordering), id. at 61-62 (Visualization), id. at 62 (Time Sharing); Levy Decl. Ex. II, at 54-55 (Visualization); Levy Decl. Ex. JJ, at 47-48 (Inductive Reasoning, Deductive Reasoning).) These difficulties seemed to stem from the fact that panel members were not informed by the test-maker of the meaning of the abilities. (See Int. 56.1 ¶ 95.)

The parties' expert submissions highlight these deficiencies in the City's job analysis. First, although the City bears the burden to show a proper job analysis, the City's Bobko-Schemmer Report does little to satisfy this burden. The Bobko-Schemmer Report simply summarizes the test construction process, as described in the Test Development Report, and provides a few parenthetical comments about the steps taken. It states that the Test Development Report:

- updated important task statements from [Dr. Landy's] prior task list;
- collected firefighter importance ratings of tasks and cognitive abilities, as well as links between these two domains (a process that is widely used in industrial-organizational psychology to provide a basis for demonstrating content validity);
- collected the above information using targeted interviews/observations, a focus group, and a job analysis survey completed by a sample of 192 firefighters which included ethnic/racial minority groups and females;
- used an ability taxonomy that is largely based on work by Fleishman (the Fleishman taxonomy provides a consistent framework for researchers to examine incumbent and expert perceptions regarding the ability demand of jobs);

- used nine of these abilities to write [questions] for the written exam (nine abilities also formed the basis for [questions] in the earlier [Dr.] Landy . . . written exam); and
- used and trained panels of incumbent firefighters as [question] writers and [question] reviewers, with attention to diversity of these panels. (Using incumbents as preliminary [question] writers has several potential advantages. It should help ensure that the [question] content is consistent with the firefighter job. Also, the language and reading level of the [question] text will tend to be consistent with that in the incumbent population.)

(Bobko-Schemmer Report 28.)[26] According to Dr. Bobko and Dr. Schemmer, "the cognitively-based written exams were developed following standard job analytic and test development procedures—thus speaking to their job relatedness." (Id.) Notably, however, the Bobko-Schemmer Report does not address the deficiencies in the linking process.

By contrast, Plaintiffs' expert reports provide specific reasons to doubt the validity of the City's job analysis process. According to the expert opinion of each of Plaintiffs' experts, "the flaws in the [City's] job analysis were fatal to the validity of the exams." (Int. 56.1 ¶ 101.) This is not simply a difference in opinion among experts—Plaintiffs' experts set out specific problems with the City's job analysis that the City's experts never address.

For example, the Jones-Hough Report includes four pages of criticism about the reliability of judgments made by the Linking Panel. (See Jones-Hough Report 26-30.) It states that "[t]he linking panel judgments on which [the City's] examination development plan was based appear to have been done without sufficient understanding on the part of the linking panel members." (Jones-Hough Report 35-36.) According to Dr. Jones and Dr. Hough, it was "[t]roubling in this stage of the project [that] information regarding the degree to which the 12

[26] Although the Bobko-Schemmer Report states that the question-writing panels were created with "attention to diversity," the Test Development Report actually states that, although a "female and [H]ispanic firefighter" were requested for the panel, "the agency was unable to comply with this request." (Test Development Report 12.) There was one black male firefighter on the panel. (Id.)

firefighters involved in producing the final written examination specification did not understand and perform their assignment." (Id. at 26.) The Jones-Hough Report summarizes this critique as follows:

> [T]hough linking panel members appear to have experienced problems in performing the task they were presented, their judgments were used to determine the number of questions that would be used to assess each of the cognitive abilities measured by the new written examination. In our professional opinion, this represents a fatal flaw in the information used to determine the test development plan for Written Exam 7029.

(Id. at 29; see also Goldstein Report 15 (criticizing the work of the linking panel).) The City never addresses these deficiencies in the Linking Panel judgments, and based on the undisputed evidence, a fact finder could not conclude that abilities and tasks were properly matched.[27]

Another deficiency identified Plaintiffs experts relates to the specific tasks and abilities selected for measurement. In his expert report submitted for Plaintiffs, Dr. Goldstein opines that the City inappropriately retained tasks and abilities in its job analysis that did not meet a "Day One" standard—in other words, the City tested for tasks and abilities that could be learned on the job. (Goldstein Report 12.) Citing the EEOC Guidelines, Dr. Goldstein explained that a "content valid test should measure work behaviors, activities, and/or worker [knowledge, skills, abilities or characteristics] that are important for the performance of the job and are needed at entry, rather than learned on the job." (Id. at 12 (citing 29 C.F.R. § 1607.14C(1) ("Content validity is also not an appropriate strategy when the selection procedure involves knowledge,

[27] On this point, the City argues only that a change in a person's view of the importance of certain tasks to the job of firefighter does not necessarily undermine that person's prior answer, because it may simply reflect a change of perspective over time. (Def. Int. 56.1 ¶ 96.) For this unexceptional proposition, the City cites several depositions which provide support for it. (See, e.g., Fraenkel Decl. Ex. 18 ("Wiesen Dep."), at 102; Fraenkel Decl. Ex. 19, at 35.) But, this argument does not rebut Plaintiffs' evidence that various Linking Panel members' ratings were compromised by the failure to understand the definitions of the listed abilities when asked to perform the linking analysis.

skills, or abilities which an employee will be expected to learn on the job.").) As Dr. Goldstein opined, "content validity models are concerned with establishing that the content of a test reflects the content of a job—in other words, that what a candidate must do to perform well on the test corresponds to what a worker must do to perform well on the job. That critical content validity link is broken if what the worker must do to perform well on the job is learned after entering the job (and, thus, after taking the test)." (Id.) This Day One standard is also reflected in 29 C.F.R. § 1607.5F, which states that employers should "avoid making employment decisions on the basis of measures of knowledge, skills, or abilities which are normally learned in a brief orientation period, and which have an adverse impact." The City does not address its failure to comply with the EEOC regulations setting out a Day One standard.

Instead of attempting to address these deficiencies, the City resorts almost entirely to citing the work performed by Dr. Landy on Exam 0084, a predecessor to the examinations at issue in this case. The City repeatedly argues that it was reasonable for it to rely upon Dr. Landy's validation study and test plan in devising Exams 7029 and 2043. (Def. BN Mem 5 ("Exam 7029 was developed in 1999 and was based on the work done by outside consultant Dr. Frank Landy for Exam 0084.").) Although the City cites extensively to Dr. Landy's involvement in the development of Exam 0084, the cited evidence shows that his work had only limited effect on Exams 7029 and 2043.

The Test Development Report states that the task list developed by Dr. Landy was used as part of an "updated task list" that reflected new input from Morrongiello for Exam 7029. (See Test Development Report 5 ("The information derived from the interviews/observations and the task list from the previous job analysis were incorporated into an updated task list.").) At his

deposition, Morrongiello confirmed that he used Dr. Landy's work in this way, testifying that he "started out with the task list that was used on 0084 just as a starting point . . . ." (Morrongiello Dep. 478-79; see also Test Development Report 4 (referring to use of Fleishman's ability list and use of those abilities on prior exam).) Indeed, Morrongiello stated at his deposition that he had merely "referred" to Dr. Landy's report, but that he had not "read [it] in detail." (Morrongiello Dep. 440; see also id. at 478 (stating that he relied on Dr. Landy "to a degree").) There is simply no evidence presented that Dr. Landy's work played any other role in the process of developing the examinations at issue. The undisputed evidence shows that the tasks and abilities lists for Exams 7029 and 2043 used Dr. Landy's work on Exam 0084 as a starting point, nothing more.

In spite of his limited role, the City nonetheless refers repeatedly to the work of Dr. Landy in arguing for the validity of Exams 7029 and 2043. The City cites several times to Dr. Schemmer's statement, which has been stricken, that "given Dr. Landy's stature in the field it would be hard to imagine that one of Dr. Landy's studies would possess substantial defects." (Def. BN Mem. 5.) As the EEOC Guidelines explicitly provide, however, reliance on the stature of a test-maker cannot stand in for a proper showing of validity. See 29 C.F.R. § 1607.9A ("Under no circumstances will the general reputation of a test or other selection procedures, its author or its publisher, or casual reports of [its] validity be accepted in lieu of evidence of validity.").[28] The mere presence of Dr. Landy in the process of identifying tasks and abilities for

[28] The only other evidence on which the City relies is a statement from the stricken Schemmer Declaration that "Exams 7029 and 2043 are in content and substance very representative of entry level firefighter selection exams which used more rigorous methods and which were thoroughly documented." (Def. BN. Mem. 7 (quoting Schemmer Decl. ¶ 5).) Even were it to be considered by the court, this bare statement, unaccompanied by any citation, analysis, or other discussion, cannot stand in for the analysis required by Guardians. As the Second Circuit cautioned in Gulino, a court's reliance on expertise should "be tempered by the scrutiny of reason and the guidance of Congressional intent." 460 F.3d at 383 (internal quotation marks omitted). Moreover, "casual reports of [an examination's] validity [cannot] be accepted in lieu of evidence of validity." 29 C.F.R. § 1607.9A. With no basis provided by Dr. Schemmer—and no particular tests, "rigorous methods" or documentation cited or discussed—his

Exam 0084 does not allow the City to ignore the problems in its job analysis for Exams 7029 and 2043.

Moreover, the City fails to confront the fact that the Landy Report was very different from the analysis and process used to construct Exams 7029 and 2043. (See Int. BN Reply 8-10.) The Federal Government points out that "Dr. Landy's own report regarding the work he did for the City (which is labeled on its front cover a 'Draft') explicitly states that he was not able to complete a job analysis because the firefighters' union refused to cooperate, and only 217 of the 5,500 job analysis questionnaires Dr. Landy sent to FDNY firefighters were completed and returned." (USA BN Mem. 8-9 (citing Landy Report 1-2, 14).) Like the other deficiencies identified by Plaintiffs, these remain unaddressed by the City.

In sum, with respect to the first Guardians requirement, a reasonable fact finder could conclude that the City's job analysis adequately began by identifying tasks and abilities important to the job of entry-level firefighter. Yet, the undisputed evidence shows that the City nonetheless failed to establish the relationship between the tasks it identified and the abilities it sought to test, and that it failed to rely on a Day One standard in assessing what abilities should be tested. Accordingly the City's job analysis in this case is, like the showing in Guardians, of "questionable sufficiency." 630 F.2d at 96.

2. Test Construction Process

The second Guardians requirement is a proper test construction process. In analyzing this requirement, Guardians set forth two relevant points of guidance. First, Guardians explained that civil service examinations should be constructed by testing professionals. Although observing

---

opinion does not support the validity of Exams 7029 or 2043, nor does it support the more specific requirement of a suitable job analysis.

that "the law should not be designed to subsidize specialists," the Second Circuit cautioned that "employment testing is a task of sufficient difficulty to suggest that an employer dispenses with expert assistance at his peril." Id. Therefore, Guardians criticized the City for allowing police officers themselves to write test questions. Id. ("The questions were initially framed by police officers, who may have had expertise in identifying tasks involved in their job but were amateurs in the art of test construction."). Second, in Guardians the City never tested its examination questions for reliability, nor did it "perform[] the minimal sample testing to ensure that the questions were comprehensible and unambiguous." 630 F.2d at 96. The Second Circuit therefore cautioned that examination questions should be tested to ensure their reliability. Id.

In constructing Exams 7029 and 2043, the City has ignored the Second Circuit's guidance. The Test Development Report makes clear that no outside expertise was utilized to construct test questions. Instead, the City relied upon panels of firefighters to write the questions for the challenged examinations. While input from incumbent firefighters was crucial in determining what tasks firefighters do, input from testing professionals was needed to devise questions that could assess which candidates would better perform those tasks. See Guardians, 630 F.2d at 97. However thoroughly a test-maker determines the important tasks of firefighters, the resulting examination will be deficient if its questions fail to connect to those tasks or fail to identify which candidates are best equipped to perform them. The City ignored Guardians' warning the municipalities should rely on expert assistance in constructing civil service examinations. See Guardians, 630 F.2d at 96; cf. Ricci, 2009 WL 1835138, at *5 (test constructed by exam specialist); Fickling v. N.Y.S Dep't of Civil Serv., 909 F. Supp. 185, 190 (S.D.N.Y. 1995) (tests for New York State welfare eligibility examiner not constructed by

testing specialists); Cuesta v. N.Y.S. Office of Court Admin., 657 F. Supp. 1084, 1097 (S.D.N.Y. 1987) (noting that the Office of Court Administration "duly heeded" the Guardians warning to have expert assistance in test construction).

Second, the City has presented no evidence that it performed any sample testing to ensure its examinations adequately and reliably tested the nine identified abilities. Guardians, 630 F.2d at 96; Cuesta, 657 F. Supp at 1097-98 (civil service examination questions were "pilot-tested on selected sample populations to measure their difficulty, impact, and validity"). Because the City has not presented evidence of sample testing, the court is left without any confidence that Written Exams 7029 and 2043 reliably tested the abilities identified by the City's job analysis.

In sum, the City has not offered any evidence of a competent test construction process under the second Guardians requirement. Instead, the City argues the same points as it did in support of its job analysis. (Def. BN Mem. 7.) Even were these arguments sufficient to show an adequate job analysis—which they are not—they are insufficient to satisfy the separate requirement of competent test construction. Viewed together, these inadequacies in the overall test development process mirror those in Guardians—indeed, the City appears to be relying on the same practices for which it was criticized by the Second Circuit thirty years ago.

3. Direct Relationship

Guardians' third requirement is that the content of the test be directly related to the content of the job. The third requirement reflects "[t]he central requirement of Title VII" that a test be job-related. 630 F.2d at 97-98. In Guardians, the court was satisfied that the "abilities that were actually tested for . . . adequately related to most of the identified tasks." Id. at 98. The abilities tested were "filling out forms," "remembering facts," and applying "general

principles to specific fact situations." Id. The Second Circuit was satisfied that these were the abilities needed to be a police officer. Id.

Here, the City has offered evidence from which a fact finder could conclude that the abilities it attempted to test had some relationship to the job of entry-level firefighter. The nine abilities that the City intend to test on Exams 7029 and 2043 were: Written Comprehension, Written Expression, Memorization, Problem Sensitivity, Deductive Reasoning, Inductive Reasoning, Information Ordering, Spatial Orientation, and Visualization. (See Levy Decl. Ex. M, at 4, 6 (Admission ## 28, 35).) Although these are all cognitive abilities, the City has presented sufficient evidence that the nine abilities reflect, to some degree, the job of entry-level firefighter. This uncontroversial point does not appear to be disputed.

A major flaw in the City's showing, however, is identified in the expert opinion of Dr. Siskin that Exams 7029 and 2043 did not actually test those nine abilities. To reach this conclusion, Dr. Siskin conducted two statistical analyses. One analysis measured the "correlation" among test questions: this analysis presupposes that, if test questions are actually measuring the ability they are intended to measure, then questions measuring the same ability will be more highly "correlated" with each other than with questions measuring other abilities. (See Siskin II Report 5; see also Fraenkel Decl. Ex. 14 ("Cline Dep."), at 322.) Dr. Siskin measured the correlation of each of the nine abilities with itself and with each other ability for Exams 7029 and 2043. (See Siskin II Report 5-6 & tbls. 1 & 2.) This analysis revealed a pattern showing that the "[questions] intended to measure an individual cognitive ability actually tend[ed] to correlate as or more highly with [questions] intended to measure different cognitive abilities . . . ." (Id. at 6.) "Four of the nine abilities [had questions] that correlate[d] on average

more highly with [questions] intended to measure different abilities than with [questions] intended to measure the same ability." (Id. (emphases added).) For Exam 7029, "everything except the [questions] intended to measure Spatial Orientation correlate[d] most highly with Written Expression."[29] (Id.) These correlation patterns led Dr. Siskin to conclude that "the [questions] on Written Exams 7029 and 2043 do not measure nine distinct abilities, as they were designed to do," and that "the written examinations fail to measure and weight the nine ability constructs consistent with what the test developer's job analysis deemed to be relevant to performance." (Id. at 7.)

To further support his conclusion, Dr. Siskin applied a method called "factor analysis," which is "a statistical methodology that, based on the empirical data, defines an underlying structure which can explain the correlation among the [questions]." (Id.) "For the results of factor analysis to confirm the test plan, the analysis should find that [questions] group together to comprise nine or 10 factors in a manner consistent with the test plan, such that the Deductive Reasoning [questions] group together to form one factor and the [questions] intended to measure Inductive Reasoning group together to form a second factor, and so forth."[30] (Id. at 7-8.) Dr. Siskin's factor analysis showed that the data did not "factor into nine distinct factors or ability domains," but instead "seem to primarily measure a general cognitive ability (except, perhaps, Memorization), and to a much lesser extent, a second specific cognitive ability (which is

[29] According to the court's review, some of the examination questions appear to have tested for Spatial Orientation. (See Fraenkel Decl. Ex. 31 (Exam 7029 questions 26, 31, 39, 43-45, 53, 61, 66, 71, 72-74, 75; Exam 2043 questions 15-18, 34-35, 51-54, 65-66, 78-80, 81-83).)

[30] As Dr. Siskin explains, "[i]n addition to nine factors corresponding to the nine discrete abilities, a single common factor may also be expected to measure a general cognitive ability (i.e., general intelligence) that would influence all the nine discrete abilities." (Id. at 8 n.4.)

different from any defined by the test developers)."[31] (Id. at 9). According to Dr. Siskin, "[t]his result demonstrates that the purported intent of the test design (to measure and weight nine distinct cognitive ability domains) was not successful." (Id.)

Dr. Siskin's analysis undermines the City's contention that the examinations tested for nine distinct cognitive abilities, rather than for cognitive ability in general. In response, the City argues that a determination that the examinations tested cognitive ability in general is consistent with expectations. According to the City, experts recognize that, "when a factor analysis is conducted the analysis should yield only one factor," because "cognitive measures are highly intercorrelated." (Def. BN Mem. 9.) For this assertion, the City relies upon the declaration of Dr. Cline, whose expert opinion has been stricken. Even if considered, however, the cited statement from the Cline Declaration does not establish a relationship between the abilities supposedly tested and the actual content of the examinations, but only attempts to rebut analysis tending to disprove that relationship. Moreover, the Cline Declaration does not address the "correlation" analysis performed by Dr. Siskin, a separate statistical analysis showing that nine cognitive abilities were not tested.[32] In any case, apart from its criticisms of the assumptions underlying the analysis of Plaintiffs' experts, the City has presented no evidence to satisfy its burden to show the relationship between the specific abilities it attempted to test and the questions it used to test them.[33]

---

[31] According to the court's review, some of the examination questions appear to have tested explicitly for Memorization. (See Fraenkel Decl. Ex. 31 (Exam 7029 questions 1-11; Exam 2043 questions 1-11).)

[32] Other deposition testimony from Dr. Cline supports reliance on correlation analysis. (See USA BN Reply 12 (citing, inter alia, Fraenkel Decl. Ex. 16 ("Cline II Dep."), at 321-22, 325-26).)

[33] The Federal Government also argues that Dr. Siskin's correlation and factor analyses demonstrate that the City has not shown what the content of their examinations is, and that the City cannot show unknown content to be related to the job of firefighter. (USA BN Mem. 10-12.)

In spite of the City's weak showing on the third Guardians requirement, a fact finder could conclude that cognitive abilities, such as Spatial Orientation, Memorization and Reading Comprehension, were tested on the written examinations, and that some of those abilities are required of entry-level firefighters. Accordingly, a reasonable fact finder could conclude that the some of the cognitive abilities tested are relevant to the job of firefighter.

4. Representativeness Requirement

As its fourth requirement, Guardians demands that a test be a "representative sample of the content of the job." 630 F.2d at 98 (internal quotation marks omitted). This requirement has two components: "[t]he first is that the content of the test must be representative of the content of the job; the second is that the procedure, or methodology, of the test must be similar to the procedures required by the job itself." Id. In setting forth these requirements, Guardians did not require that "all the knowledges, skills, or abilities required for the job be tested for, each in its proper proportion," nor did it require that a test "simulate the actual work setting" of the job. Id. Instead, the Second Circuit stated: "it is reasonable to insist that the test measure important aspects of the job, at least those for which appropriate measurement is feasible, but not that it measure all aspects, regardless of significance, in their exact proportions." Id. at 99. An important purpose of this requirement is to "prevent either the use of some minor aspect of the job as the basis for the selection procedure or the needless elimination of some significant part of the job's requirements from the selection process entirely . . . ." Id. at 99.

In addition, the Second Circuit specifically recognized as part of the fourth requirement that "although all pencil and paper tests are dependent on reading, even if many aspects of the job are not, the reading level of the test should not be pointlessly high." Id.

i. Abilities Tested

In finding that the test in Guardians met the representative requirement, the Second Circuit was satisfied that the abilities it measured were "all significant aspects of entry-level police work." Id.[34] The evidence is starkly different in this case. In support of the content of Written Exams 7029 and 2043, the City's Bobko-Schemmer Report notes only that the job of firefighter has "cognitive demands." In particular, it identifies one of the 21 task clusters considered by the City's job analysis, "Size Up," and concludes that this task cluster "invoke[s] cognitive processes." (See Bobko-Schemmer Report 27-28.)

This conclusion is insufficient to meet the requirement of representativeness. It is not enough to state that one of the task clusters from the City's job analysis has cognitive demands. Instead, the City must provide evidence of what the important abilities of a firefighter are, and must demonstrate the extent to which those abilities were tested on its examinations. The City has not done so. Instead, the undisputed evidence shows that the City failed to test for cognitive and non-cognitive abilities that are important to the job of firefighter. This evidence shows that the cognitive abilities intended to be tested on Exams 7029 and 2043 were not the most important cognitive abilities for the job of firefighter. Moreover, it shows that non-cognitive abilities are more important to the job than cognitive abilities.

The City did not test all the important cognitive abilities of the job of firefighter. According to the City's own job analysis, eighteen cognitive abilities were deemed important to the job of firefighter, but only nine of those abilities were tested. (See Test Development Report

[34] Guardians noted that its "conclusion would have been easier to reach if the City had spelled out the relationship between the abilities that were tested for and the job behaviors that had been identified[, but that] the relationship [was] sufficiently apparent to indicate that the City was not seizing on minor aspects of the police officer's job as the basis for selection of candidates." 630 F.2d at 99.

11; see also Levy Decl. Ex. M, at 4, 6 (Admission ## 28, 35).) The two most important abilities were excluded because they involved oral abilities. (Patitucci II Dep. 131-32, 274-75; Int. 56.1 ¶ 106.) With regard to the seven other omitted abilities, Morrongiello testified that he "didn't do anything specific to determine" whether or not they were testable, and that he was "going by . . . standard operating procedure at that time in our unit . . . that these abilities" would not be tested. (Morrongiello Dep. 443.) The City has provided no explanation for why it did not test these seven cognitive abilities that its own job analysis deemed important.

Besides these cognitive abilities, the City has also recognized non-cognitive abilities as important to the job of firefighter. For example, the City has admitted the importance of: Resistance to Stress, Teamwork, Responsibility, Desire to Learn, Honesty, Cleanliness, Medical Interest, Achievement Orientation, Dependability, and Conscientiousness. (Int. 56.1 ¶¶ 114, 115; Def. Int. 56.1 ¶ 115; see also Levy Decl. Ex. CC, at 9 (Admission # 65).) The city's expert, Dr. Bobko, has acknowledged that non-cognitive abilities are important to the job of entry-level firefighter. (See Int. 56.1 ¶ 117.)

It is undisputed by the City that, "[a]ll other things being equal, an examination that measures more of the knowledge, skills, abilities or characteristics that are important for a job is expected to be more valid than an examination that measures fewer of the [knowledge, skills, abilities or characteristics] that are important for that job." (Int. 56.1 ¶ 103; see also Levy Decl. Ex. CC, at 8; Fraenkel Decl. Ex. 12 ("Bobko Dep."), at 317.) Nevertheless, the City does not explain why it ignored this principle in devising the abilities to be tested on Exams 7029 and 2043.

Plaintiffs' experts provide further support for the conclusion that cognitive and non-cognitive abilities, which were not tested on Exams 7029 and 2043, are important to the job of firefighter. For example, in their expert report, Dr. Jones and Dr. Hough remark that "job analysis studies . . . show that the firefighter position involves a host of requirements other than cognitive abilities. All were ignored by the City's job analysis study." (Jones-Hough Report 19.) The Jones-Hough Report cites two "significant, nationwide government-funded projects" that "provide a long established history of the importance of several non-cognitive characteristics needed for effective performance of firefighter tasks." (Id. at 24.) According to the Jones-Hough Report, the first project was performed in approximately 1975 under contract from the U.S. Office of Personnel Management ("OPM") in order to, inter alia, "provide job information that local fire departments could use in developing their own firefighter selection procedures." (Id. at 22 (internal quotation mark omitted).) The nationwide study found that 11 of the 20 "required firefighter abilities and characteristics" were non-cognitive, and that only four were cognitive. (Id. (internal quotation marks omitted).)

Similarly, as the Jones-Hough Report indicates, non-cognitive abilities are identified by the Department of Labor's Occupational Information Network ("O*NET"), which "provides comprehensive job descriptive information including information about knowledge, skills, abilities and work styles."[35] (Id. at 23.) According to Dr. Jones and Dr. Hough, this source identifies the following non-cognitive abilities as an important part of the municipal firefighter job: Dependability, Cooperation, Coordination, Concern for Others, Service Orientation, Social

[35] According to the website for the Employment and Training Administration in the Department of Labor, "[t]he O*NET database is a comprehensive source of descriptors, with ratings of importance, level, relevance or extent, for more than 900 occupations that are key to our economy." See U.S. Dep't of Labor, Employment & Training Admin., "O*Net – beyond information – intelligence," available at http://www.doleta.gov/programs/ONet/ (last visited on July 21, 2009).

Orientation, Initiative, Persistence, Attention to Detail, Self Control, Stress Tolerance, and Active Listening. (Id. at 24.)

In his expert report, Dr. Wiesen opined that the City's job analysis omitted important abilities and characteristics. (Wiesen Report 31-36.) In doing so, Dr. Wiesen pointed out that various non-cognitive abilities have been recognized as important to the job of firefighter and that, prior to the creation of Exam 7029, those abilities were considered testable. The Wiesen Report references the development of Written Exam 6019 (id. at 35), the O*NET categories (id. at 34), and a study published by the U.S. Civil Service Commission in 1977 with the goal of "identify[ing] the major tasks performed by entry-level firefighters across the United States."[36] (Levy Decl. Ex. PP, at VUL 03160; see Wiesen Report 32.) As explained in the Wiesen Report, the Civil Service Commission study shows that the five most important attributes of the job of firefighter are non-cognitive, and that, of the 20 abilities included in the Civil Service Commission's "Weighting Plan" for firefighter examinations, only four were cognitive. (Wiesen Report 32.)

Dr. Wiesen also stated in his expert report that "[j]ob analysis studies done in various jurisdictions nationwide prior to the time of preparation of Exam 7029 have shown the importance of non-cognitive abilities for successful job performance as a Firefighter," and cites studies. (Id. at 34.) For example, the Wiesen Report cites a study performed by OPM in 1980 which showed the importance of "mechanical ability" to the job of firefighter. (Id. at 35.) The Wiesen Report further states that the City "did not test mechanical ability, a cognitive ability that

[36] The Intervenors have produced a copy of the 1977 report of the U.S. Civil Service Commission, entitled "Job Analysis of the Entry Level Firefighter Position." (See Levy Decl. Ex. PP.) The Executive Summary states that that the "nationwide job analysis of the entry-level firefighter position" included a sample of 109 fire departments representing cities from 20,000 to more than 2.5 million people. (Id. at VUL 03160.)

could have been tested easily and which has been reported to be a valid cognitive ability for predicting job performance of Firefighters, and for which tests have been available for many years." (Id.; see also Fraenkel Decl. Ex. 8 ("Schemmer Dep."), at 263 (noting that "mechanical" ability could have been tested at the time of the challenged examinations).)

Similarly, the Goldstein Report criticizes the City's decision to test only cognitive abilities in creating Exams 7029 and 2043. In that expert report, Dr. Goldstein pointed out that neither Morrongiello (who developed Exam 7029) nor Johnston (who developed Exam 2043) considered testing non-cognitive abilities. He then opined that "[t]here is no good scientific or professional basis given for the decision not to conduct a more comprehensive job analysis and test development process." (Goldstein Report 16.)

Of course, Guardians did not require that "all the knowledges, skills, or abilities required for the job be tested for, each in its proper proportion." 630 F.2d at 98. But the failure of Exams 7029 and 2043 to test for a considerable number of abilities that are undisputedly important to the job of firefighter, even according to the City's own evidence, raises significant doubts about the representativeness of those examinations. Cf. Fickling, 909 F. Supp at 192 ("The test seized upon relatively minor aspects of the Eligibility Examiner job, such as reading comprehension and arithmetic and ignored others.").

Some of these exclusions are understandable. For example, the City argues "that it would not have been feasible for the City of New York in 1999 and 2002 to have administered structured interviews or oral comprehension tests as part of the entry level firefighter examination." (Def. BN Mem. 10.) This may be true, but Plaintiffs are not challenging merely the failure of the City to utilize tests for oral comprehension, or tests based on structured

interviews. Instead, they challenge the City's failure to test for important cognitive and non-cognitive abilities, while administering a test for less relevant abilities as a threshold for entry into the Academy.

The City appears to recognize this, and argues only that no better testing was available at the time Exams 7029 and 2043 were administered. Citing the Cline Declaration, the City argues that it would not have been feasible, prior to the construction of Exam 6019 in 2007, to test for the additional abilities that she added for that examination. (See Def. BN Mem. 10-11 (citing Cline Decl. ¶¶ 10-11).) These statements have been stricken and, in any case, do not support the City's assertion. Instead, the Cline Declaration would actually support a conclusion that tests of non-cognitive abilities were available prior to the development of Exam 6019. In it, Dr. Cline states that, in late 2001, Dr. Cline used "Situational Judgment Exercises," eventually used for Exam 6019, in her development of a different civil service examination for the City. (Cline Decl. ¶ 5; see also Cline Dep. 509-12 (indicating that Situational Judgment Exercises have been considered valid since a study in 2001).) Moreover, Dr. Cline indicates that Perceptual Speed could have been tested on Exams 7029 and 2043. (Cline Decl. ¶ 7.) The City does not explain why this ability was not tested.

Moreover, even assuming that certain abilities tested on Exam 6019 could not have been tested before that examination was developed, this does not absolve the City's failure to test any non-cognitive abilities on Exams 7029 and 2043. There is undisputed evidence that tests for non-cognitive abilities have been available for decades. For example, as the Intervenors point out, the DCAS Examiner for Exam 6019 testified at his deposition that non-cognitive tests based on "biodata" have been available since at least the 1980s. (See Fraenkel Decl. Ex. 25

("Alexander Dep."), at 350-57.) The Intervenors also point to the U.S. Civil Service Commission's study on the job of firefighter in 1977, which includes an appendix listing "studies showing [the] empirical validity" of testing for various non-cognitive abilities. (See Levy Decl. Ex. PP, at 578-82.) Indeed, the City's own expert, Dr. Schemmer, testified that written examinations evaluating non-cognitive abilities were available in 1999 and 2002 when Exams 7029 and 2043 were administered. (See Schemmer Dep. 292-98; see also Levy Decl. Ex. RR.) Based on this evidence, the City has no excuse for its failure to test important cognitive and non-cognitive abilities.

ii. Reading Level

Aside from the City's failure to test additional abilities on Written Exams 7029 and 2043, the City has also failed to satisfy a separate aspect of Guardians' fourth requirement: an appropriate reading level. Guardians explicitly warned that the reading level of a written examination "should not be pointlessly high." Id. at 99. The City has not heeded this warning. It is undisputed that an analysis of the reading level was never conducted before the examinations were administered, and the City does not now present an analysis to demonstrate an appropriate reading level. (See Int. 56.1 ¶ 128; Def. Int. 56.1 ¶ 128.)

On the other hand, Plaintiffs' experts conducted an analysis showing that the reading level of the two challenged examinations was too high. Cf. Ricci, 2009 WL 1835138, at *5 (noting that officers' test was constructed at a reading level below the tenth grade). Dr. Wiesen analyzed the reading level of Written Examinations 7029 and 2043, and concluded that the reading level was "above the 12th grade and 'was too high for the job of Firefighter.'" (Int. 56.1 ¶ 127 (quoting Wiesen Report 58).) To reach this conclusion, Dr. Wiesen calculated the reading

level of each examination as a whole, as well as for individual questions, relying on a test called the "SMOG" formula.[37] His reading level analysis shows that the average reading grade levels for 7029 and 2043 questions were 12.3 and 12.8, respectively. (Wiesen Report 59, 60.) Moreover, fifty of the questions on Exam 7029 and fifty-six of the questions on Exam 2043 had a reading level above the 12th grade. (Id. at 59.) By comparison, Dr. Wiesen observed that the reading level under the SMOG formula for another widely used firefighter's examination was at the 10th grade reading level. (Id.) Dr. Wiesen also stated that, on both examinations, black candidates left more of the last ten questions blank than white candidates. (Id. at 74.) He found this disparity statistically significant and opined that it might be attributed to the unnecessarily high reading level. (Id.)[38]

Dr. Wiesen also found the reading level to be inappropriate because reading conditions during the examinations differed from reading conditions at the Academy. Cf. Guardians, 630 F.3d at 99 (noting that "risks of using a written test were substantially minimized" because "[t]he reading level necessary to understand the questions was in some cases equal to, but generally well below, the training materials used in the Police Academy"). Although Dr. Wiesen did not asses the reading level of the Academy's training materials, he did contrast the nature of the

[37] According to one journal, "[t]he SMOG formula . . . uses the number of polysyllabic (>=3 syllables) words per sentence to estimate the minimal grade reading level required for full (100%) comprehension of educational materials . . . ." Richard Rogers et al., The Language of Miranda Warnings in American Jurisdictions: A Replication and Vocabulary Analysis, 32 Law & Hum. Behav. 124, 127 (2008).

[38] In response, the City cites the Schemmer Declaration. The court has stricken that declaration. Even were the court to consider it, however, it is not helpful to the City. Dr. Schemmer opines that the reading analysis method used by Dr. Wiesen "was not intended to examine multiple choice test," but was intended for "standard extended prose passages." (Schemmer Decl. ¶ 17.) But Dr. Wiesen directly addressed this point by observing that the majority of questions on each examination contained more than 100 words, and that, overall, Written Examination 7029 contained 11,844 words and Written Exam 2043 contained 11,517 words. The court's review of the examinations reveals numerous questions that contain sizable text passages. (See, e.g., Fraenkel Decl. Ex. 31 (Exam 7029 questions 12-14, 16-18, 24, 36-37, 46-47, 50-52, 57, 60, 63, 65, 80; Exam 2043 questions 19-22, 24-27, 28-30, 32, 37-38, 40-41, 67-69, 83).)

Academy training environment with the nature of a testing environment. He observed that, "[w]hen written material is used in the training of Firefighters, both in the Fire Academy and in the fire station, the firefighters are allowed and even encouraged to ask questions of supervisors, trainers or other colleagues." (Wiesen Report 58 (citing deposition testimony).) Conditions for taking a written multiple-choice test are different from the Academy because "talking (and asking questions) is forbidden." (Id. at 59.) Accordingly, "the reading level requirements of written test questions should be below the reading level of material Firefighters may be given to read at the Fire Academy or in the fire station where assistance is readily available." (Id.)

Considered together, the undisputed evidence relating to the fourth Guardians requirement—that the content of Exams 7029 and 2043 be representative of the job of firefighter—is extremely weak. From its failure to test various cognitive and non-cognitive abilities to its failure to show that the examinations had an appropriate reading level, the City has not provided sufficient evidence that Exams 7029 and 2043 were a "representative sample of the content of the job." Guardians, 630 F.2d at 98. The City's showing in this case is well below what it showed in Guardians thirty years ago.

5. Scoring System

In Guardians, the court observed that the City's evidence of content and representativeness (under requirements 3 and 4) had been adequate, while its evidence of proper job analysis and test construction (under requirements 1 and 2) had departed "in some significant respect even from reasonably attainable requirements." 630 F.2d at 99. Following these conclusions, the court went on to find that, "even if the construction of the exam [had] passe[d] muster, the way in which it was used to distinguish among candidates seriously departs from the

[fifth Guardians requirement] . . . and defeats any claim of validity for a testing process the produces disparate racial results." Id. at 99-100. The City's failures on the fifth requirement have the same dispositive force in this case.

i. Cutoff Scores

In addressing the use of cutoff scores, Guardians observed that "[n]o matter how valid the exam, it is the cutoff score that ultimately determines whether a person passes or fails. A cutoff score unrelated to job performance may well lead to rejection of applicants who were fully capable of performing the job." 630 F.2d at 105. This common-sense principle is embodied in the EEOC Guidelines, which provide that a cutoff score "should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force." Id. (quoting 29 C.F.R. § 1607.5(H)). As the Second Circuit held in Guardians, "[w]hen a cutoff score unrelated to job performance produces disparate racial results, Title VII is violated." Id.

Accordingly, "there should generally be some independent basis for choosing the cutoff." Id. For example, "the employer might establish a valid cutoff score by using a professional estimate of the requisite ability levels, or, at the very least, by analyzing the test results to locate a logical 'break-point' in the distribution of scores." Id. In Guardians, the Second Circuit criticized the City for simply choosing "as many candidates as it needed, and then set[ting] the cutoff score so that the remaining candidates would fail." Id.

This method had a particularly problematic effect in Guardians because of reliability concerns about the examination. Id. at 105-06. As stated by the court, the "reliability" of an examination is "the extent to which the exam would produce consistent results if applicants

repeatedly took it or similar tests." Id. at 101. "[I]f an exam lacks reliability to such an extent that results would be significantly inconsistent if the same applicants were to take it again, that is an important indication that the test is not especially useful in measuring their abilities." Id. Reliability concerns are especially significant because of what testing experts call "the error of measurement." Id. at 102. For any given test, a candidate's score on successive versions of that test will inevitably differ based upon nothing more than chance, since small differences in scores might not reflect meaningful differences in a test taker's aptitude. See id. Accordingly, grades within a certain range of one another should "theoretically be treated as equivalent" in order to account for the likelihood that small differences have resulted from chance. Id. at 102-03. Based on the error of measurement, the Guardians court expressed concern that the City's selection of a cutoff score had led to a high number of "mistaken passes and failures," because so many scores near the selected cutoff score likely resulted from chance differences. Id. at 105-06. The court therefore concluded that, "[w]hen an exam produces disparate racial results, a cutoff score requires adequate justification and cannot be used at a point where its unreliability has such an extensive impact as occurred in this case." Id. at 106.

The evidence presented by the City indicates that it relied on inappropriate grounds in selecting the cutoff scores for Exams 7029 and 2043. For Exam 7029, the cutoff score was based merely on the number of entry-level firefighter job openings expected by the FDNY. (See Fraenkel Decl. Ex. 9 ("Wachter Dep."), at 74-75, 181-82; Patitucci Dep. 91-93.) In making this selection, the City ignored Guardians' warning that the City should not "simply chose "as many candidates as it needed, and then set the cutoff score so that the remaining candidates would fail." 630 F.2d at 105. For Exam 2043, the City based its cutoff score on the "default" set by the

City's civil service rules. (See Int. 56.1 ¶ 63.) This choice of cutoff score ignored Guardians warning against relying solely on a civil service default score. See id. at 104-05 (rejecting City's reliance on civil service law requirements for setting rank-ordering); see also id. ("Title VII explicitly relieves employers from any duty to observe a state hiring provision which purports to require or permit any discriminatory employment practice.") (internal quotation marks omitted); Fickling, 909 F. Supp at 192 ("[Defendants] merely rely upon [civil service rules] which set 70% of the total possible score as the passpoint.").

More importantly, the City has presented no evidence that its chosen cutoff scores bear any relationship to the necessary qualifications for the job of entry-level firefighter. See Fickling, 909 F. Supp at 193 ("[Defendants] have not offered any evidence, however, that the passpoint was either a logical 'break-point' in the distribution of scores or that it corresponded to the ability level required by the job."); see also Lanning v. Southeastern Pa. Transp. Auth., 181 F.3d 478, 489 (3d Cir. 1999) ("[I]n order to show the business necessity of a discriminatory cutoff score an employer must demonstrate that its cutoff measures the minimum qualifications necessary for successful performance of the job in question."). The City has conceded that the cutoff scores were not selected in order to "measure the minimum level of the tested skills, abilities or other characteristics necessary for successful performance of the job of entry-level firefighter in the FDNY." (Int. 56.1 ¶ 59; see also Levy Decl. Ex. X, at 96-97.) Nor were they selected based upon a validity study or the job analysis performed by the City. (See Wachter Dep. 85; Patitucci Dep. 93.)

The expert evidence only further undermines the City's reliance on the scores. The City's Bobko-Schemmer Report does not address the cutoff scores, which Dr. Bobko explicitly

recognized at his deposition. (See Bobko Dep. 319 ("Q: In the parts of your report where you talk about job relatedness and business necessity, you didn't discuss the pass/fail cutoff scores the City used on written exam 7029 and 2043, correct? / A: Correct."). In fact, Dr. Bobko answered "No" when asked whether the his report was "sufficient to establish" that the City's use of Exams 7029 and 2043 as a pass/fail screening device was job-related and consistent with business necessity. (See Fraenkel Decl. Ex. 15 ("Bobko II Dep."), at 179-80.)

By contrast, Plaintiffs' expert evidence suggests that the cutoff scores improperly screened out candidates. As set forth in his expert report, Dr. Siskin conducted an analysis of all candidates who sat for both Exam 7029 and 2043. Because the construction and content of the examinations were largely the same, Dr. Siskin's expectation was that candidates who took both examinations should have done roughly as well on each. To determine whether this was the case, Dr. Siskin posited a hypothetical situation in which the cutoff score for each examination was 70, and analyzed whether, based on that score, candidates sitting for both examinations would have passed each examination. Dr. Siskin found discrepancies in the resulting data:

> Among those who took both written examinations, 54.8 percent (17 out of 31) of those who failed Written Exam 2043 scored 70 or above on Written Exam 7029. Of those who scored below 70 on the Written Exam 7029, 75.9 percent (44 out of 58) passed Written Exam 2043. Thus, of those failing either examination at a 70 percent cutoff score, 81.3 percent (61 out of 75) failed one written examination but passed the other. Since these figures reflect a single candidate taking different administrations of what is purportedly the same test, the high degree of inconsistency affirmatively highlights the fact that the cut-off scores on the written examinations do not reliably predict the presence or absence of the minimum cognitive skills and abilities to do the job.

(Siskin II Report 19; see also Jones-Hough Report 39-40.) Dr. Siskin stated that, "[c]omparing the rankings of all 2,667 candidates who took both examinations, we find the typical change in an individual's rank between Written Exam 7029 and Written Exam 2043 was 458 positions (a

change of about 20 percent up or down) on the list." (Siskin II Report 24-25.) Some candidates moved up as many as two thousand places in rank, and some moved down as many as two thousand places.[39] (Id. at 24.)

In its defense, the City points to the deposition testimony of Dr. Jones to argue that "one method for establishing a cut-off score, which has business justification, is based on hiring needs." (Def. BN Mem. 11.) The testimony the City cites does not support its position:

> Q: What would be wrong with using the . . . anticipated need of hires or future potential candidates as the basis for establishing the cut-off score?
>
> A: Well, I think it can be a consideration but using it to set the cut-off score specifically and explicitly, bears no relationship to the standard that people in my profession like to see, and that is one of job-relatedness. To say we need to have X number of people available is not a[] statement of job-relatedness at all. It's a statement that says: Here is how many people we want to pass. We don't know whether those people passing all are either . . . qualified or . . . not qualified. Perhaps some of them who were rejected would have been qualified. All we know is that we set the bar so that we get enough people through the funnel, and that's not what I would consider accepted practice in our discipline.
>
> Q: Would you consider that a business requirement?
>
> A: There is some business justification for taking into account how many people are needed, but that's a long distance from using that information to justify how you proceed.

(Fraenkel Decl. Ex. 13 ("Jones Dep."), at 80-82.) The City cites only to the last sentence, but neglects to mention Dr. Jones' clear, primary point that relying on hiring needs alone is insufficient to constitute business justification for setting a cutoff score.

The City also argues that its cutoff scores were not unnecessarily high. It argues that there is no need to reduce a cutoff score to place people on an eligibility list if those people will

[39] The City argues that this discrepancy might be due simply to the passage of time. (See Def. BN Mem. 14-15.) In his expert report, however, Dr. Siskin considered that some of the changes might be due to the lapse of time or change in the candidates' skills or knowledge, but opined that the large changes in rank were not likely due to such factors. (Siskin II Report 25 & n.37.)

never be reached. For this basic proposition, it points to the testimony of Dr. Cline, stating that "there's no reason to . . . hire more people than you can use," and that a cutoff score might reflect the fact that "even if you put people on the list, they're not going to get hired." (Cline Dep. 458-59.) The City contends that this proposition is supported by 29 C.F.R. § 1607.5H, which, its asserts, "allow[s] for cutoff scores higher than minimal proficiency if candidates scoring below that score 'have little or no chance of being selected for employment.'" (Def. BN Mem. 12 (quoting 29 C.F.R. § 1607.5H).)

Yet, the City fails to cite the entire regulation. The regulation begins by stating that "[w]here cutoff scores are used, they should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force." 29 C.F.R. § 1607.5H. As discussed, the City has not attempted to show that its cutoff scores bear any relationship to "normal expectation of acceptable proficiency" for the job of entry-level firefighter. The regulation goes on to state that:

> Where applicants are ranked on the basis of properly validated selection procedures and those applicants scoring below a higher cutoff score than appropriate in light of such expectations have little or no chance of being selected for employment, the higher cutoff score may be appropriate, but the degree of adverse impact should be considered.

Id. (emphasis added). As the Federal Government points out, the City has produced no evidence that candidates who scored below the cutoff score had "little or no chance of being selected." (USA BN Mem. 15.) The evidence suggests the opposite. The cutoff score did not just determine who would be excluded from the eligibility list, but also prevented applicants from taking the PPT. Because they were never given the opportunity to take the PPT, those applicants were deprived of the opportunity to place ahead of passing candidates on the list. (See Mar. 19,

2009 Tr. 56 ("If you are not ranking on the same thing you are setting the cutoff score on, then you can't say they had little or no chance of being hired.").) Because those who failed the written examination were not permitted to take the PPT, those who failed the written portion may have been selected based on the combination of their written and PPT scores. As the Government puts it, "a candidate who scored 83.529 [a failing score] on Written Exam 7029 and would have scored 100 on the PPT (if the City had allowed him to take it) would have a combined score of 83.652, much higher than a candidate who scored 84.705 [a passing score] and received a score of 75 on the PPT." (USA BN Mem. at 15-16 (citing Int. 56.1 ¶ 65).) This analysis undermines the City's position.

Contrary to the City's position, 29 C.F.R. § 1607.5H does not permit reliance on a "higher cut-off score" whenever candidates who are ranked at the bottom of a list would not be hired. Although the regulation allows an employer to set a higher cutoff score when an eligibility list resulting from a lower score might result in candidates not being reached, it also requires that, in order to do so, there must first be a cutoff score that is "reasonable and consistent with normal expectations of acceptable proficiency within the work force." See 29 C.F.R. § 1607.5H. If certain candidates on an eligibility list have "little or no chance of being hired" based on a proficiency-based cutoff score, then it may be permissible to use a higher score so that they are not needlessly waiting on a list that will never reach them. Even then, the "degree of adverse impact should be considered." Cf. M.O.C.H.A., 2009 WL 604898, at *17 (noting that testing professional considered degree of adverse impact in setting cutoff score). The regulation does not, as the City suggests, give a municipality carte blanche to raise or set a cutoff score according only to hiring needs.

The evidence convincingly demonstrates that impermissible grounds were relied upon by the City in selecting cutoff scores for the two challenged examinations. This is all the more troubling because of the City's exceedingly weak justification for the construction and content of Exams 7029 and 2043. Without any showing that "independent" job-related grounds played any role whatsoever in the City's cutoff decisions, there is no basis to proceed to a trial on the City's pass/fail uses of Exams 7029 and 2043. For these reasons, summary judgment must be granted to Plaintiffs with respect to the City's use of Written Examination 7029 as a pass/fail screening device with a cutoff score of 84.705, and the City's use of Written Examination 2043 as a pass/fail screening device with a cutoff score of 70.

ii. Rank-Ordering

The discussion in Guardians of rank-ordering revolved around two basic principles. First, the court explained that "content validity is not an all or nothing matter; it comes in degrees." Id. at 100. As the court explained:

> [R]ank-ordering makes such a refined use of the test's basic power to distinguish between those who are qualified to perform the job and those who are not. If a test is content valid, it may be reasonable to infer that the test scores makes some useful gross distinctions between candidates. Candidates with high scores may well be expected to perform the job better than candidates with low scores. And it may even be that within some range of scores, some incremental improvements in scores show some positive correlation with improvements in job performance. But neither of these propositions provides confidence for inferring that one-point increments among those who took [the examination] are a valid basis for making job-related hiring decisions, especially in the range of scores between 94 and 100. . . . [That is, a] test may have enough validity for making gross distinctions between those qualified and unqualified for a job, yet may be totally inadequate to yield passing grades that show positive correlation with job performance.

Id. In other words, "[r]ank-ordering satisfies a felt need for objectivity, but it does not necessarily select better job performers." Id; see also Cuesta, 657 F. Supp. at 1101 ("[T]o use

single point differentials on an examination of such limited scope to wholly determine hiring priority over-inflates the importance of those abilities measured."). This reasoning recognizes that the use of rank-ordering based upon the results of an examination requires strong justification. See Lindemann 159 ("Courts generally require employers to present specific and well-documented justification for . . . rank-order selection and, accordingly, frequently reject rank-ordering selection practices as inadequately validated or unjustifiably adverse.").

Second, Guardians identified problems in rank-ordering where the reliability of an examination has been called into question. According to the court, "[w]ithout some substantial demonstration of reliability it is wholly unwarranted to make hiring decisions, with a disparate racial impact, for thousands of applicants that turn on one-point distinctions among their passing grades." Guardians, 630 F.2d at 101. The court highlighted two concerns. The first concern was the "quality of the exam questions." Id. The more skillfully that test questions have been formulated, the more certain a court can be that they reliably test what they aim to test, and that scores do not "vary because of extraneous factors such as test administration." Id. at 101-02. The second concern regarded the "error of measurement," which is discussed above. Id. at 102.

Based on these potential problems surrounding the reliability of exam results, the court identified "several ways to increase the justification for rank-ordering sufficiently to use it." Id. at 103-04. First, the test-maker "can conduct a job analysis and construct the test with a high degree of adherence to Guideline requirements." Id. at 104. To do so, "there must be a substantial demonstration of job relatedness and representativeness to show a sound basis for making rank-ordering hiring decisions." Id. Second, "the test-maker can achieve an adequate degree of reliability by careful design of the exam so that the questions will yield a satisfactory

degree of consistent results." Id. To do so, the test-maker should "pre-test the exam by successive applications to an appropriate sample or at least analyze the result of split-half correlations."[40] Id. A test-maker can then correct "[i]nconsistencies revealed by these methods." Id. Third, a test-maker can simply "acknowledge [his or her] inability to justify rank-ordering and resort to random selection from within either the entire group that achieves a properly determined passing score, or some segment of the passing group shown to be appropriate." Id.

Whether or not an employer chooses one of these methods, the court observed, it "cannot use rank-ordering not shown to be job-related when test scores produce a disparate racial impact." Id. This is precisely what the City has done here. The City has not presented any evidence that its use of rank-ordering is job-related. As with its cutoff scores, the City's expert, Dr. Bobko, stated that his expert report did not establish the validity of using rank-ordering based on Exams 7029 and 2043. (See Bobko II Dep. 179-80.) Moreover, Plaintiffs' expert evidence shows that the City's rank-ordering led to problematic results. For example, the Siskin II Report shows how insignificant differences in candidates' scores on the written examinations could result in sizable differences in their ranking. For example, Dr. Siskin looked at those who scored 100 and just below on Written Exam 7029, and opined that:

> a candidate who scored 100 on the PPT but got one question wrong on Written Exam 7029 would have fallen from a maximum rank of 1 [had he or she not answered the question incorrectly] to a maximum rank of 440. A second wrong answer on Written Exam 7029 (i.e., a score of 98.82) would have dropped the candidate's rank to a maximum of 1,146; four wrong answers to Written Exam 7029 (i.e., a score of 95.294) would have dropped the candidate's maximum rank to 2,454 on the list. Given the standard error of measurement of Written Exam 7029 (i.e., 2.64) a wrong answer to four questions on Written Exam 7029 is

[40] Split-half correlation is a method of "dividing each component of the test into equal halves and observing how consistent were an individual's score on each half." Id. at 102.

> within the range of normal variation in scores due to chance (and may not represent any true difference between individuals).

(Siskin II Report 23.) Similarly, for Exam 2043, "four incorrect answers rather than a perfect score (which would constitute a drop from a score of 100 to a score of 95.294 on Written Exam 2043) with a 100 on the PPT, would have dropped a candidate['s] rank from a maximum rank of 1 to a maximum rank of 1,713." (Id.) In other words, for both examinations, statistically insignificant differences in test performance amounted to significant differences in ranking.

In its defense, the City argues that the caselaw recognizes the availability of rank-ordering when an examination is "valid." (Def. BN Mem. 13 ("If an examination is valid, rank ordering and selection of candidates is permissible.").) The City goes on to argue that, since "Plaintiffs have not and cannot demonstrate that the examinations are invalid," they cannot show that rank-ordering is inappropriate. (Id.) The City maintains that the validity and reliability of Exams 2043 and 7029 render their results sufficient for rank-ordering.

The City argues from faulty premises. First, Plaintiffs are not obligated to demonstrate the "invalidity" of the examinations. Instead, the City bears the burden to show that its hiring practices are job-related. See Gulino, 460 F.3d at 382 (citing 42 U.S.C. § 2000e-2(k)(1)(A)(i)); see also 42 U.S.C. § 2000e(m). Second, "validity" in the context of employment examinations is not "an all or nothing matter, it comes in degrees." 630 F.2d at 100. This is because "[a] test might have enough validity for making gross distinctions between those qualified and unqualified for a job, yet may be totally inadequate to yield passing grades that show positive correlation with job performance." Id. These principles from Guardians contradict the City's assertion that the limited steps it took to ensure test validity allow it to make fine distinctions

among candidates based on small differences in test scores. Although the rank-ordering use of the examinations may have satisfied "a felt need for objectivity," the City was wrong to "make hiring decisions, with a disparate racial impact, for thousands of applicants" based on fine distinctions in scores on the challenged examinations. Id.

The City also argues that, because all examinations have some margin of error, "there will always be a risk of choosing between candidates whose scores, though different, may fall within the margin of error." (Def. BN Mem. 15.) But the Second Circuit was not blind to these concerns when it ruled in Guardians that the City's use of rank-ordering was impermissible. The Second Circuit explicitly recognized the high burden its approach placed upon employers, explaining: "[i]f test scores produce disparate racial results, an employer who wants to use rank-ordering of the scores for hiring decisions faces a substantial task in demonstrating that rank-ordering is sufficiently justified to be used." 630 F.2d at 103 (emphasis added). The court then identified several options available to an employer, including either: (1) a substantial demonstration of adequate job analysis and test construction, (2) an adequate showing of reliability of the examination, supported by pre-testing or split-half testing of the examination, or (3) random ranking of qualified candidates. Id. at 103-04. The City's submissions fail to address any of these options.

Finally, the City relies upon the deposition testimony of firefighters, as lay witnesses, to support its use of rank-ordering. These firefighters "believe that people higher on the list are likely to perform better than those ranked lower down." (Def. BN Mem. 16.) However, their lay opinion testimony does not support the rank-ordering used by the City. For example, the City points to a firefighter's deposition testimony that, "if you're able to write as well on the written

test and the score is high on the physical test, then yeah, you're probably going to do better than someone at the end of the list[] than someone who isn't as physically capable or doesn't have as much deductive reasoning." (Fraenkel Decl. Ex. 27, at 108.) The testimony continued, however:

> Q: Have you actually seen this reflected in your experience of working with firefighters, that those who came off the top—
>
> A: I don't pay attention to list numbers. In my experience, I don't know. I don't know. A new guy walks in the door tomorrow, I don't know if he's on the top of the list or the bottom of the list.
>
> . . . Once you step in the door after school is over, it's nonessential where you were on the list . . . .

(Id. at 108-10.) This witness did not even know the rankings of other firefighters. Another witness similarly testified that, although he was "told" that those higher on the list performed better in "training school," he had "never experienced" any difference in performance, and that "[n]o one in the firehouse . . . knew anybody's test number." (Fraenkel Decl. Ex. 28, at 72.)[41]

The cited testimony does not support the "refined use" of the Exams 7029 and 2043 that is required for rank-ordering. At most, it reflects a general acceptance of the written examinations among firefighters, or the ungrounded assumption that those at the top of an eligibility list must perform better than those at the bottom. Such "nonempirical or anecdotal accounts" cannot substitute for actual evidence of validity. See 29 C.F.R. § 1607.9A. Needless to say, these anecdotal accounts do not support the high degree of correlation between

[41] The City's third witness provided an explanation for this assumption that is totally unrelated to the aptitude of test takers. A firefighter captain testified that, in general, those at the top of a hiring list perform better than those at the bottom of a hiring list. (See Fraenkel Decl. Ex. 29, at 63-65.) But, importantly, he did not attribute this difference in performance to differences in aptitude. Instead, he testified that candidates lower down on an eligibility list have to wait longer to become firefighters and, "when you get down towards the end you're getting older people" who are more "set in their ways" and harder to "mold" into firefighters. (Id. at 108-09.) In other words, this witness attributed the difference in performance to the delay caused by being ranked lower, rather than attributing a lower ranking to a difference in capability.

examination score and job performance that is needed to justify rank-ordering. See Guardians, 630 F.2d at 100.

Based on the evidence presented, no reasonable fact finder could conclude that the claimed reliability or validity of Exams 7029 and 2043 sufficiently supports rank-ordering of candidates. The challenged examinations have produced a disparate impact upon minority candidates, and the City has made an inadequate showing that the tests contained appropriate content and were properly constructed pursuant to the EEOC Guidelines and Guardians. Under these circumstances, the City has not justified its ranking of thousands of candidates for the job of entry-level firefighter. The court must grant summary judgment for Plaintiffs with respect to the City's rank-order processing and selection of candidates from the Written Exams 7029 and 2043 eligibility lists.

V. CONCLUSION

There can be no doubt that the job of firefighter is crucial to the health, safety and security of the City's communities. Our firefighters routinely display selflessness and bravery in fulfilling their critical responsibilities. In the hopes of joining their ranks, tens of thousands of applicants—including thousands of minority applicants—sat for Exams 7029 and 2043. These examinations, however, have had an undeniable adverse impact on black and Hispanic candidates, excluding them from positions as entry-level firefighters, and closing the doors of opportunity for public service to large segments of the City's population.

There is no disputed issue of material fact on which to proceed to trial in the disparate impact case. Plaintiffs have established, based on undisputed evidence, that the City's uses of Exams 7029 and 2043 have adversely affected black and Hispanic candidates. They have shown

systemic disparities of statistical and practical significance in the pass/fail rates and eligibility list rankings of those minority candidates. These unlawful practices barred over a thousand additional black and Hispanic applicants from consideration for appointment as FDNY firefighters, and unfairly delayed the appointment of hundreds of black and Hispanic firefighters. Accordingly, Plaintiffs have established a prima facie case of disparate impact on account of race and national origin in violation of Title VII of the Civil Rights Act of 1964.

In its defense, the City has failed to raise a triable issue that this disparate impact was the result of business necessity. The City has failed to demonstrate a sufficient relationship between the tasks of a firefighter and the abilities it intended to test on Exams 7029 and 2043. It has failed to take measures to ensure the reliability of those examinations; it has failed to take steps to ensure that that the reading level of the examinations was appropriate; it has failed to test for various recognized important abilities of a firefighter; it has failed to test for abilities needed upon entry into the Academy, rather than abilities to be learned on the job; it has failed to retain testing professionals to devise the examination questions; and it has failed to demonstrate that the examinations it administered actually tested the abilities it intended to test. Compounding these failings, the City has imposed arbitrary pass/fail scores, unrelated to the qualifications for the job of entry-level firefighter, and has constructed eligibility lists based on distinctions in test scores that are unrelated to corresponding differences in the qualifications of firefighter candidates. Following the Second Circuit's holding in Guardians, the court concludes that the City improperly relied upon these poorly constructed examinations in the face of a disparate impact upon minority candidates. The undisputed evidence shows that the City cannot defeat summary judgment by showing disputed fact issues as to its defense.

Accordingly, the court GRANTS Plaintiffs' and Intervenors' Motions for Summary Judgment. Plaintiffs have established disparate impact liability and an appropriate remedy must now be considered by the court. Intervenors' disparate treatment case also remains pending. The court shall now proceed with these aspects of the case.

SO ORDERED.

/s/ Nicholas G. Garaufis
NICHOLAS G. GARAUFIS
United States District Judge

Dated: Brooklyn, New York
July 22, 2009