823 F.3d 102
United States Court of Appeals, First Circuit.

Pedro LOPEZ, individually and on behalf of a class of individuals similarly situated; Abel Cano, individually and on behalf of a class of individuals similarly situated; Kevin Sledge, individually and on behalf of a class of individuals similarly situated; Charles De Jesús, individually and on behalf of a class of individuals similarly situatuated; Richard Brooks, individually and on behalf of a class of individuals similarly situated; Massachusetts Hispanic Law Enforcement Association, individually and on behalf of a class of individuals similarly situated; Robert Alvarez, individually and on behalf of a class of individuals similarly situated; Spencer Tatum, individually and on behalf of a class of individuals similarly situated; Shumeand Benfold, individually and on behalf of a class of individuals similarly situated; Angela Williams–Mitchell, individually and on behalf of a class of individuals similarly situated; Gwendolyn Brown, individually and on behalf of a class of individuals similarly situated; Lynette Praileau, individually and on behalf of a class of individuals similarly situated; Tyrone Smith, individually and on behalf of a class of individuals similarly situated; Eddy Chrispin, individually and on behalf of a class of individuals similarly situated; David E. Melvin, individually and on behalf of a class of individuals similarly situated; Steven Morgan, individually and on behalf of a class of individuals similarly situated; William E. Iraolo, individually and on behalf of a class of individuals similarly situated; José Lozano, individually and on behalf of a class of individuals similarly situated; Courtney A. Powell, individually and on behalf of a class of individuals similarly situated; James L. Brown, individually and on behalf of a class of individuals similarly situated; George Cardoza, individually and on behalf of a class of individuals similarly situated; Larry Ellison, individually and on behalf of a class of individuals similarly situated; David Singletary, individually and on behalf of a class of individuals similarly situated; Charisse Brittle Powell, individually and on behalf of a class of individuals similarly situated; Cathenia D. Cooper–Paterson, individually and on behalf of a class of individuals similarly situated; Molwyn Shaw, individually and on behalf of a class of individuals similarly situated; Lamont Anderson, individually and on behalf of a class of individuals similarly situated; Gloria Kinkead, individually and on behalf of a class of individuals similarly situated; Kenneth Gaines, individually and on behalf of a class of individuals similarly situated; Murphy Gregory, individually and on behalf of a class of individuals similarly situated; Julian Turner, individually and on behalf of a class of individuals similarly situated; Neva Grice, individually and on behalf of a class of individuals similarly situated; Delores E. Facey, individually and on behalf of a class of individuals similarly situated; Lisa Venus, individually and on behalf of a class of individuals similarly situated; Rodney O. Best, individually and on behalf of a class of individuals similarly situated; Karen Vandyke, individually and on behalf of a class of individuals similarly situated; Robert C. Young, individually and on behalf of a class of individuals similarly situated; Royline Lamb, individually and on behalf of a class of individuals similarly situated; Lynn Davis, individually and on behalf of a class of individuals similarly situated; James A. Jackson, individually and on behalf of a individuals similarly situated; Louis Rosario, Jr., individually and on behalf of a class of individuals similarly situated; Obed Almeyda, individually and on behalf of a class of individuals similarly situated; Devon Williams, individually and on behalf of a class of individuals similarly situated; Julio M. Toledo, individually and on behalf of a class of individuals similarly situated, Plaintiffs, Appellants,

v.

Marisol Nobrega, individually and on behalf of a class of individuals similarly situated, Plaintiff,

v.

CITY OF LAWRENCE, MASSACHUSETTS; City of Methuen, Massachusetts; John Michael Sullivan, in his capacity as Mayor of the City of Lawrence, Massachusetts; William Manzi, III, in his capacity as Mayor of the City of Methuen, Massachusetts; City of Lowell, Massachusetts; City of Worcester, Massachusetts; Appointing Authority for the City of Lowell, Massachusetts; Michael O'Brien, in his capacity as City Manager of the City of Worcester, Massachusetts; City of Boston, Massachusetts; City of Springfield, Massachusetts; Domenic J. Sarno, Jr., in his capacity as Mayor of the City of Springfield, Massachusetts; Massachusetts Bay Transportation Authority; Daniel Grabauskas, in his capacity as General Manager, Massachusetts Bay Transportation Authority; Board of Trustees of the Massachusetts Bay Transportation Authority, Defendants, Appellees,

William F. Martin, in his capacity as Mayor of the

City of Lowell, Massachusetts; Konstantina B. Lukes, in her capacity as Mayor of the City of Worcester, Massachusetts; Commonwealth of Massachusetts; Paul Dietl, in his capacity as Personnel Administrator for the Commonwealth of Massachusetts, Defendants.

No. 14–1952
|
May 18, 2016.

Synopsis
Background: Black and Hispanic police officers who were not selected for promotion to sergeant brought Title VII action against city, alleging that the test used for selecting officers for promotion, consisting of a written examination and an education and experience rating followed by a rank-order selection, resulted in an unjustified disparate impact based on race. The United States District Court for the District of Massachusetts, George A. O'Toole, Jr., J., entered judgment for city after a bench trial. Officers appealed.

Holdings: The Court of Appeals, Kayatta, Circuit Judge, held that:

test was a valid selection tool, and

officers failed to meet their burden of putting forward a specific less discriminatory alternative to the test.

Affirmed.

Torruella, Circuit Judge, filed a dissenting opinion.

Procedural Posture(s): On Appeal.

Attorneys and Law Firms

***106** Harold L. Lichten and Stephen S. Churchill, with whom Benjamin Weber, Lichten & Liss–Riordan, P.C., and Fair Work, P.C., were on brief, for appellants.

Bonnie I. Robin–Vergeer, Attorney, Department of Justice, Civil Rights Division, Appellate Section, with whom Sharon M. McGowan, Attorney, Civil Rights Division, Vanita Gupta, Acting Assistant Attorney General, P. David López, General Counsel, and Carolyn L. Wheeler, Acting Associate General Counsel, Appellate Services, Equal Employment Opportunity Commission, were on brief for amicus the United States of America.

Kay H. Hodge, with whom John M. Simon, Geoffrey R. Bok, Stoneman, Chandler & Miller LLP, Susan M. Weise, Attorney, City of Boston Law Department, and Lisa Skehill Maki, Attorney, City of Boston Law Department, were on brief, for appellee City of Boston, Massachusetts.

James F. Kavanaugh, Jr., with whom Christopher K. Sweeney, and Conn Kavanaugh Rosenthal Peisch & Ford, LLP, were on brief, for appellees Massachusetts Bay Transportation Authority, Daniel Grabauskas, and the Board of Trustees of the Massachusetts Bay Transportation Authority.

Rachel M. Brown, Assistant City Solicitor, City of Lowell Law Department, with whom Christine Patricia O'Connor, City Solicitor, City of Lowell Law Department, was on brief for appellees City of Lowell, Massachusetts, and Appointing Authority for the City of Lowell, Massachusetts.

Tim D. Norris, with whom Joshua R. Coleman, and Collins, Loughran & Peloquin, P.C., were on brief, for appellees City of Worcester, Massachusetts, Michael O'Brien, City Manager of Worcester, and Konstantina B. Lukes, Mayor of the City of Worcester.

Anthony I. Wilson, Associate City Solicitor, City of Springfield Law Department, with whom Edward M. Pikula, City Solicitor, and John T. Liebel, Associate City Solicitor, were on brief, for appellees City of Springfield, Massachusetts, and Mayor Domenic J. Sarno, Jr.

Raquel D. Ruano, Attorney, Office of the City Attorney, City of Lawrence, Massachusetts, and Charles D. Boddy, Jr., Attorney, Office of the City Attorney, City of Lawrence, Massachusetts, on brief for appellees City of Lawrence, Massachusetts, and Mayor John Michael Sullivan.

Kerry Regan Jenness, Attorney, Office of the City Solicitor, City of Methuen, on brief for appellees City of Methuen, Massachusetts, and Mayor William M. Manzi, III.

Michael L. Foreman, Civil Rights Appellate Clinic, Dickinson School of Law, Pennsylvania State University, on amicus brief of National Urban League and the National Association for the Advancement of Colored People.

Gary Klein, Kevin Costello, Corinne Reed, Klein Kavanagh Costello, LLP, Mark S. Brodin, Professor, Boston College Law School, and Ray McClain, Director, Employment Discrimination Project, Lawyers' Committee for Civil Rights Under Law, on amicus brief

of Massachusetts Association of Minority Law Enforcement Officers, New England Area Conference of ***107** the National Association for the Advancement of Colored People, Urban League of Eastern Massachusetts, and Professor Mark S. Brodin.

Christopher L. Brown, Christopher J. Petrini, and Petrini & Associates, P.C., on amicus brief of International Municipal Lawyers Association, Massachusetts Municipal Lawyers Association, Massachusetts Municipal Association, National Public Employer Labor Relations Association, Massachusetts Chiefs of Police Association, Inc., and Fire Chiefs Association of Massachusetts, Inc.

Before TORRUELLA, LYNCH, and KAYATTA, Circuit Judges.

Opinion

KAYATTA, Circuit Judge.

In selecting police officers for promotion to the position of sergeant in 2005 and 2008, the City of Boston and several other Massachusetts communities and state employers adapted a test developed by a Massachusetts state agency ("HRD")[1] charged under state law with creating a selection tool that "fairly test[s] the knowledge, skills and abilities which can be practically and reliably measured and which are actually required" by the job in question. Mass. Gen. Laws ch. 31, § 16. There is no claim in this case that defendants intentionally selected the test in order to disadvantage any group of applicants. To the contrary, the evidence is that the test was the product of a long-running effort to eliminate the use of race or other improper considerations in public employment decisions.

The percentage of Black and Hispanic applicants selected for promotion using the results of this test nevertheless fell significantly below the percentage of Caucasian applicants selected. Some of those Black and Hispanic applicants who were not selected for promotion sued, claiming that the use of the test resulted in an unjustified "disparate impact" in violation of Title VII notwithstanding the absence of any intent to discriminate on the basis of race. 42 U.S.C. § 2000e–2(k)(1)(A)(i). After an eighteen-day bench trial, the district court determined, among other things, that the use of the test did have a disparate impact on promotions in the City of Boston, but that the test was a valid selection tool that helped the City select sergeants based on merit. *Lopez v. City of Lawrence*, No. 07–11693–GAO, 2014 U.S. Dist. LEXIS 124139, at *37, *60–62 (D.Mass. Sept. 5, 2014). The court further found that the plaintiffs failed to prove that there was an alternative selection tool that was available, that was as (or more) valid than the test used, and that would have resulted in the promotion of a higher percentage of Black and Hispanic officers. *Id.* at *60–79. Finding that the district court applied the correct rules of law and that its factual findings were not clearly erroneous, we affirm.

I. Background

The plaintiffs in this suit (the "Officers") sought promotion in the police departments operated by the Massachusetts municipalities or state agencies sued in this case. *Id.* at *7–8. All parties agree that affirmance of the judgment in favor of Boston would result in affirmance of the judgment in favor of the other defendants as well, so we focus our discussion for simplicity's sake on the evidence concerning Boston. Because this is an appeal of fact-finding and application of law to fact following a trial on the merits, we describe ***108** the facts in a manner that assumes conflicting evidence was resolved in favor of the prevailing party unless there is particular reason to do otherwise. *Wainwright Bank & Tr. Co. v. Boulos*, 89 F.3d 17, 19 (1st Cir.1996) ("We summarize the facts in the light most favorable to the verdict-winner [ ], consistent with record support.").

A. Development of the Exams Over Time

In 1971, Congress noted that the United States Commission on Civil Rights ("USCCR") found racial discrimination in municipal employment "more pervasive than in the private sector." H.R.Rep. No. 92–238, at 17 (1971). According to the USCCR, nepotism and political patronage helped perpetuate pre-existing racial hierarchies. U.S. Comm'n on Civil Rights, *For All the People, By All the People: A Report on Equal Opportunity in State and Local Government Employment*, 63–65, 119 (1969), *reprinted in* 118 Cong. Rec. 1817 (1972). Police and fire departments served as particularly extreme examples of this practice. *See, e.g.*, Wesley MacNeil Oliver, *The Neglected History of Criminal Procedure, 1850–1940*, 62 Rutgers L.Rev. 447, 473 (2010) ("Officers who delivered payments to their superiors were practically assured of retention and even promotion, regardless of their transgressions."); Nirej S. Sekhon, *Redistributive Policing*, 101 J.Crim. L. & Criminology 1171, 1191 (2011) ("Police departments were prime sources of patronage jobs.").

Boston's police department was no exception: As far back as the nineteenth century, a subjective hiring scheme that hinged on an applicant's perceived political influence and the hiring officer's subjective familiarity with the candidate (or the candidate's last name) was seen as the primary culprit behind a corrupt, inept, and racially exclusive police force. *See, e.g.*, George H. McCaffrey, *Boston Police Department*, 2 J. Am. Inst.Crim. L. & Criminology 672, 672 (1912) ("This system worked very unsatisfactorily, however, because places on the police force were invariably bestowed as a reward for partisan activity.").

At both the state and local levels, Massachusetts officials eventually gravitated toward competitive exams as a tool to accomplish an important public policy of moving away from nepotism, patronage, and racism in the hiring and promoting of police. *Boston Chapter, N.A.A.C.P., Inc. v. Beecher*, 504 F.2d 1017, 1022 (1st Cir.1974) ("[C]ivil service tests were instituted to replace the evils of a subjective hiring process ...."); *see generally* League of Women Voters of Mass., *The Merit System in Massachusetts: A Study of Public Personnel Administration in the Commonwealth* 3–5 (1961). At the statewide level, this movement resulted in legislation and regulations aimed at ensuring that employees in civil service positions are "recruit[ed], select[ed] and advanc[ed] ... on the basis of their relative ability, knowledge and skills" and "without regard to political affiliation, race, color, age, national origin, sex, marital status, handicap, or religion." Mass. Gen. Laws ch. 31, § 1.

B. The 2005 and 2008 Exams

Born of these purposes and shaped by decades of Title VII litigation,[2] the examinations at issue in this case allowed no room for the subjective grading of applications. The total score of a test-taker who sat for the promotional examination in ***109** 2005 or 2008 was determined by two components: an 80–question written examination scored on a 100–point scale and an "education and experience" ("E & E") rating, also scored on a 100–point scale. The written examination counted for 80% of an applicant's final score and the E & E rating comprised the remaining 20%. Applicants needed an overall score of seventy to be considered for promotion. On top of the raw score from these two components, Massachusetts law affords special consideration for certain military veterans, *id.* § 26, and individuals who have long records of service with the state, *id.* § 59.

The subject matter tested on the 2005 and 2008 examinations can be traced back to a 1991 "validation study" or "job analysis report" performed by the state agency responsible for compiling the exam.[3] *See* 29 C.F.R. § 1607.14 (technical requirements for a content validity study under the Uniform Guidelines on Employee Selection Procedures); *see also Watson v. Fort Worth Bank & Tr.*, 487 U.S. 977, 991, 108 S.Ct. 2777, 101 L.Ed.2d 827 (1988) (opinion of O'Connor, J.) ("Standardized tests and criteria ... can often be justified through formal 'validation studies,' which seek to determine whether discrete selection criteria predict actual on-the-job performance.").

That 1991 report was prepared by the Massachusetts Department of Personnel Administration ("DPA"), the predecessor to HRD. In preparing the report, DPA surveyed police officers in thirty-four jurisdictions nationwide, issuing a questionnaire that sought to ascertain the kinds of "knowledge[ ], skills, abilities and personnel characteristics" that police officers across the country deemed critical to the performance of a police sergeant's responsibilities. The report's authors distilled the initial results from this survey and their own knowledge regarding professional best practices into a list of critical police supervisory traits. They then distributed this list in a second survey to high-ranking police officers in Massachusetts, who were asked to rank these traits according to how important they felt each was to a Massachusetts police sergeant's performance of her duties. DPA further refined the ranking of key skills and traits through focused small-group discussions with police sergeants and conducted a "testability analysis" of which skills could likely be measured through the written examination or the E & E component. In 2000, HRD engaged outside consultants to refresh the findings of the 1991 examination through a process similar to, though less thorough than, DPA's approach in 1991.

The written question and answer component of the examination consisted of multiple choice questions that covered many topic areas, including the rules governing custodial interrogation, juvenile issues, community policing, and firearm issues, to name a few.[4] The text of individual questions was often closely drawn from the text of materials identified in a reading list provided by the Boston Police Department ***110** ("BPD") to test-takers in advance of the exams.

For example, one question on the 2008 promotional exam asked applicants to accurately complete the following statement:

> According to [a criminal investigations textbook on the reading list], a warrantless search and seizure is

acceptable:

A. after stopping a vehicle for a traffic violation and writing a citation.

B. after obtaining the consent of the person, regardless of whether obtained voluntarily or nonvoluntarily.

C. when possible loss or destruction of evidence exists.

D. when a quick search of the trunk of a motor vehicle is desired.

In addition to completing the question and answer component of the examination, applicants listed on the E & E rating sheet their relevant work experience, their degrees and certifications in certain areas, their teaching experience, and any licenses they held.[5] Points were assigned based on the listed education and experience. For example, applicants could receive up to fifteen points in recognition of their educational attainment, with an associate's degree providing up to three points and a doctorate providing up to twelve.

After collecting and scoring the exams, HRD provided the municipalities with a list of passing test-takers eligible for promotion, ranked in order of their test scores. Mass. Gen. Laws ch. 31, § 25. Each of the municipal defendants in this case selected candidates in strict rank order based on the list they received from HRD.[6]

Because many officers achieved at least the minimum passing score of seventy and there were relatively few openings for promotion to sergeant, all of those who were promoted scored well above the minimum in both 2005 and 2008. In 2005, 9 of the 224 Black and Hispanic candidates who took the exam were promoted, whereas 57 of the 401 other candidates were promoted. In 2008, 1 of the 213 Black and Hispanic test-takers was promoted, whereas 25 of the 291 other candidates were promoted. The average scores for those who the statisticians called "minority test takers" fell below the average scores for the "non-minority test takers" by 6.4 points in 2005 and 6.6 points in 2008.

II. Analysis

We recently described in another suit against Boston the elements of a disparate impact claim. *Jones v. City of Boston*, 752 F.3d 38, 46, 54 (1st Cir.2014). In a nutshell, litigation of such a claim in a case challenging hiring or promotion decisions focuses on three questions: Do the plaintiffs show by competent evidence that the employer is utilizing an employment practice ***111** that causes a disparate impact on the basis of race; If so, does the employer show that the challenged employment practice creating this disparate result is nevertheless job-related for the position in question and consistent with business necessity; If so, do the plaintiffs show that the employer has refused to adopt an alternative practice that equally or better serves the employer's legitimate business needs, yet has a lesser disparate impact? *Id.* To prevail, plaintiffs require a "yes" answer to the first question, and either a "no" to the second question or a "yes" to the third question. *See id.*

In this case, all parties agree that, using competent statistical analysis, the Officers have proven that Boston's use of the challenged exam in 2005 and 2008 did indeed have a marked disparate impact because the selection rates of Black and Hispanic officers for promotion to sergeant were so much lower than the selection rates of the other officers that we can fairly exclude random chance as the explanation for the difference.[7]

A. Validity

The focus of the trial thus turned to the second question: Did Boston use a "practice [that was] 'job related for the position in question and consistent with business necessity.' " *Ricci v. DeStefano*, 557 U.S. 557, 578, 129 S.Ct. 2658, 174 L.Ed.2d 490 (2009) (quoting 42 U.S.C. § 2000e–2(k)(1)(A)(i)). The parties agree that, in the context of hiring or promotion decisions, this inquiry trains on whether the selection practice—here, the use of the exam—is "valid." In simple terms, a selection practice is valid if it materially enhances the employer's ability to pick individuals who are more likely to perform better than those not picked.

In this case, Boston sought to carry its burden of proving the validity of its exams by demonstrating what the Equal Employment Opportunity Commission ("EEOC") refers to as "content validity" under the Uniform Guidelines on Employee Selection Procedures ("Guidelines"). *See* 29 C.F.R. § 1607.16(D). The parties agree generally that establishing content validity in this context requires a "showing that the content of the selection procedure is representative of important aspects of performance on the job for which the candidates are to be evaluated." *Id.* § 1607.5(B). This means that the "behavior(s) demonstrated in the selection procedure are a representative sample of the behavior(s) of the job in question or that the selection procedure provides a representative sample of the work product of the job." *Id.* § 1607.14(C)(4). Work

behavior(s) selected for measurement should either be "critical work behavior(s)" or "important work behavior(s) constituting most of the job," or both. *Id.* § 1607.14(C)(2).

Much of the evidence at trial and many of the arguments in the briefs on appeal focus on the Guidelines' technical testing standards. The Officers' briefs treat the Guidelines as if they were inflexible and binding legal standards that must be rigorously applied in ascertaining whether an employment selection device significantly advances the employer's business needs. For two reasons, this is not so.

First, "[b]ecause 'Congress, in enacting Title VII, did not confer upon the EEOC authority to promulgate rules and ***112** regulations,' the agency's guidelines receive weight only to the extent of their 'power to persuade.' " *Jones*, 752 F.3d at 50 n. 14 (quoting *E.E.O.C. v. Arabian Am. Oil Co.*, 499 U.S. 244, 257, 111 S.Ct. 1227, 113 L.Ed.2d 274 (1991)). In *Jones* itself, we rejected the Guidelines' view that plaintiffs need carry the burden of proving "practical significance" in order to establish a prima facie case of disparate impact. *Id.* at 50–53. And in *Ricci*, the Supreme Court's most recent disparate impact decision, the Court found New Haven's firefighter promotional exam job-related without mentioning the Guidelines' extensive technical criteria for assessing job-relatedness. *See Ricci*, 557 U.S. at 587–89, 129 S.Ct. 2658.

Second, even on their own terms, the Guidelines poorly serve the controlling role assigned to them by the Officers in challenging the district court's findings. The Guidelines quite understandably provide no quantitative measure for drawing the line between "representative," 29 C.F.R. § 1607.5(B), and nonrepresentative samples of job performance and behaviors. Rather, the Guidelines point to the qualitative understandings of these concepts generally accepted by professionals who evaluate "standardized tests and other selection procedures, such as those described in the Standards for Educational and Psychological Tests prepared by a joint committee of the American Psychological Association." *Id.* § 1607.5(C).

All that being said, Boston did not shy away from seeking to show that its process for selecting new police sergeants in 2005 and 2008 satisfied the technical requirements of the Guidelines. To make such a showing, the City presented the testimony of Dr. James Outtz. Outtz is an industrial organizational psychologist with twenty years of experience testing and measuring employee selection systems. He has served as a consultant to numerous American municipalities and federal agencies and has assisted in the development of employment selection devices used by many public employers. Outtz has published approximately twenty academic publications in the field of industrial organizational psychology. He has worked for both plaintiffs and defendants in challenges to the validity of exams. In *Ricci*, for example, Outtz co-authored an amicus brief brought on behalf of industrial psychologists arguing that the New Haven Fire Department promotional examinations for captain and lieutenant were flawed and invalid. *See* Br. of Industrial–Organizational Psychologists as Amici Curiae at 3, *Ricci*, 557 U.S. 557 (Nos.07–1428, 08–328), 2009 WL 796281, at *3.

Outtz reviewed the development, application, substance, and results of the exams at issue in this case. He opined that the exams were based on job analyses that validly identified the critical skills used by actual police sergeants and that the tests covered a "representative sample" of the content of the job. *Id.* § 1607.14(C)(4). In support of this conclusion, Outtz testified that the two job validity reports relied on in composing the 2005 and 2008 exams were not too stale to serve as useful starting points for the test-makers, nor were the reports otherwise infirm from a technical standpoint. While the reports—particularly the 1991 report—were somewhat dated, Outtz testified that the relative stability of a police sergeant's responsibilities over time, combined with the presence of the 2000 study, cured any defect introduced by the passage of time.[8]

***113** Outtz went on to opine that the written question and answer portion of the exam, standing alone, nevertheless did not pass muster under the Guidelines because it fell short of testing a "representative sample" of the key qualities and attributes that were identified by the two validation reports. *Id.* In Outtz's opinion, however, the addition of the E & E component effectively pushed the selection device as a whole across the finish line to show validity. It did this, according to Outtz, because the level and extent of work and educational experience and accomplishments listed by each applicant served as a useful, if imperfect, proxy for the kinds of qualities that were deemed to be important to a sergeant's daily responsibilities, yet were insufficiently tested by the examination's question and answer component. Outtz recognized that the gain in validity from the E & E component was, on its own, only marginal or "incremental." As the Officers stress, many of the attributes for which the E & E assigned points (e.g., previous service as a police officer) were shared by all or most applicants. Thus, while the E & E score range for the 2005 exam was 0–100, the actual score distribution approximated 40–94. And when weighted to provide only 20% of the combined final score, it accounted for a range of only about 5% to 7% of a candidate's total score.[9] Nevertheless, we cannot see how a rational factfinder could ignore the impact of the E & E, small or not, in evaluating the exam overall.

Outtz concluded that "at the end of the day" the combined "package" of the written examination and the E & E as administered tested a "representative sample" of the key supervisory skills identified by the 1991 and 2000 reports and was "minimally valid" or "acceptable" under the Guidelines. *Id.* He testified that the representativeness of the skills tested by the two components and the linkage of these skills to the validation reports were in line with what was contemplated by the Guidelines' technical standards for constructing a content-valid selection device. *See id.* §§ 1607.5(B); 1607.14(C)(4).

This is not to say that Outtz's testimony trumpeted a wholehearted endorsement of the scheme used by Boston to identify candidates for promotion. He agreed with the Officers that the validity of the Boston examination could have been improved, perhaps by incorporating a "well-developed assessment center" to evaluate an officer's interpersonal skills through observed social interaction, or some kind of device for measuring an applicant's oral communication skills. Outtz was clear that his opinion solely concerned the selection device's compliance with his profession's minimum standards as translated into the EEOC's Guidelines.

The Officers challenged Outtz's conclusions on cross-examination, arguing that his testimony fell short of the mark in several respects that we will discuss, and presented the contrary opinions of their own expert, Dr. James Wiesen. Altogether, the trial testimony of these competing experts consumed the better part of nine days of the eighteen-day trial.

The district court judge who listened to these experts testify concluded that Outtz was correct: "After consideration of the evidence as a whole, I find and conclude that Dr. Outtz's opinion rests on adequate ***114** grounds and is therefore correct: the exams in question were minimally valid." *Lopez,* 2014 U.S. Dist. LEXIS 124139, at *60–61. Supporting this conclusion, the court found that the examinations tested a representative sample of skills that were identified by the 1991 and 2000 reports, which were themselves valid under the Guidelines. *Id.* at *61. Finding that Boston employed valid examinations that reliably achieved the City's stated business need, the court ruled in Boston's favor. *Id.* at *78.

On appeal, the Officers now ask us to set aside the district court's finding that the 2005 and 2008 exams were valid. In considering such a request, we ask whether the district court applied the correct legal standards and whether the record contained sufficient support for its findings. *See, e.g., Beecher,* 504 F.2d at 1022 (affirming a finding of invalidity as "supported by the record"). Since our decision in *Beecher,* all circuit courts that have addressed the question have reviewed a district court's determination that a selection method was or was not valid for clear error. *See M.O.C.H.A. Soc'y, Inc. v. City of Buffalo,* 689 F.3d 263, 275 (2d Cir.2012); *Ass'n of Mex.–Am. Educators v. California,* 231 F.3d 572, 584–85 (9th Cir.2000) (en banc) ("The question whether a test has been validated properly is primarily a factual question, which depends on underlying factual determinations regarding the content and reliability of the validation studies that a defendant utilized."); *Melendez v. Ill. Bell Tel. Co.,* 79 F.3d 661, 669 (7th Cir.1996); *Hamer v. City of Atlanta,* 872 F.2d 1521, 1526 (11th Cir.1989); *Bernard v. Gulf Oil Corp.,* 890 F.2d 735, 739 (5th Cir.1989).

With this standard in mind, we consider the Officers' critique of the district court's reliance on Outtz's opinion in finding the examinations valid. Repeatedly, the Officers suggest that Outtz's own characterization of the exams as "minimally valid" should render his opinion legally insufficient to carry the City's burden. Implicitly, the Officers ask us to read "minimally valid" as meaning, in effect, "not valid enough." Read in context, however, Outtz was plainly testifying that he found the exams to be valid, albeit not by much. Indeed, elsewhere in his testimony he made clear that the exams were "valid" and, in his view, complied with the technical requirements of the Guidelines.

Moving more aptly from debating adverbs to discussing the law, the Officers (with the support of the United States as amicus curiae) argue that the district court misconstrued the law in finding Outtz's testimony sufficient. Specifically, they say that the district court did not reach its finding of content validity in accord with the Guidelines' statement that evidence of an exam's content validity should "consist of data showing that the content of the selection procedure is representative of important aspects of performance on the job for which the candidates are to be evaluated." 29 C.F.R. § 1607.5(B). Instead, argue the United States and the Officers, the district court simply counted up the knowledge, skills and abilities ("KSAs") called for by the job without qualitatively considering their importance.

It is true that the district court observed that "more than half of the KSAs identified as pertinent to the job of sergeant were tested," and that "this was sufficient to meet the 'representative sample' requirement of the Uniform Guidelines." *Lopez,* 2014 U.S. Dist. LEXIS 124139 at *54 (quoting 29 C.F.R. § 1607.14(C)(4)). The district court made this statement, though, only after first expressly citing the Guidelines standard, *id.* at *15–17, and after undertaking an examination of the tested KSAs

"to ensure that there is a link ***115** between the selection procedure and the critical KSAs necessary for successful performance of the job," *id.* at *16. The court then made clear that its examination of the manner in which the exams tested KSAs trained on "the knowledge, skills and abilities which can be practically and reliably measured and which are actually required to perform the primary or dominant duties of the position for which the examination is held." *Id.* at *50–51 (quoting Mass. Gen. Laws ch. 31, § 16). The district court also cited to testimony establishing that knowledge of the constitutional and regulatory law applicable to police work is "critical to a police sergeant's ability to effectively perform as a supervisor" and to evidence that a written job knowledge test is "[a]n effective way" of testing whether a candidate possesses such critical knowledge. *Id.* at *51–52. Similarly, the district court found that the 1991 job analysis upon which the exams were based identified "the frequent and critical tasks and duties" and the "important [KSAs] required at the time of appointment."[10] *Id.* at *52. In short, in referring to the KSAs identified as pertinent to the job of sergeant, the district court was plainly referring to the "critical" and "important" KSAs that it found to have been identified in the job analysis upon which the exams were predicated.

The district court's qualitative focus on the importance of the factors that the exam tested was further highlighted by the court's agreement with Outtz that the written job knowledge portion of the test was not alone valid "because it could not measure some skills and abilities (as distinguished from knowledge) essential to the position." *Id.* at *60. After then agreeing with Outtz that the E & E component of the exams adequately, albeit minimally, filled in this gap, the district court expressly found that the exams "were based on job analyses that considered the important tasks necessary to the successful performance of the job." *Id.* at *61. The district court's opinion as a whole thus makes clear that the court trained its focus on critical and important knowledge, skills, and abilities called for by the job, and it did not clearly err by finding that a test that measured a large percentage of such critical and important KSAs was a test that was sufficiently "representative of important aspects of performance on the job." 29 C.F.R. § 1607.5(B).[11] Our conclusion to this effect ***116** finds further support in the absence of any quantitative measure of "representativeness" provided by the law. Rather, the relevant aim of the law, when a disparate impact occurs, is to ensure that the practice causing that impact serves an important need of the employer, in which case it can be used unless there is another way to meet that need with lesser disparate impact. We cannot see how it is an error of law to find that an exam that helps determine whether an applicant possesses a large number of critical and necessary attributes for a job serves an important need of the employer.

The Officers and the United States also contend that our 1974 opinion in *Beecher,* 504 F.2d 1017, mandates our reversal of this conclusion. Their reliance on *Beecher* fits this case awkwardly because of the burdens we have already detailed. In *Beecher,* the central question was whether the record supported the district court's finding of fact that a hiring exam given to would-be firefighters was not valid. *See id.* at 1022–23. To affirm, we needed only to find that the record did not compel a contrary finding. *Id.* at 1022. Here, by contrast, the Officers ask us to find that this record compels a finding contrary to that reached by the district court.

The Officers and the United States nevertheless seem to find much significance in one analogy we drew in *Beecher.* In assessing an exam for the position of firefighter, we compared knowledge of firefighting terminology to knowledge of baseball vocabulary possessed by a potential recruit for the Boston Red Sox "who could not bat, pitch or catch." *Id.* at 1023. Here, in reviewing an exam for the supervisory position of sergeant, the more apt baseball analogy would be the hiring of a coach, who must certainly have an extensive knowledge of the rules that must be followed by those being managed. At trial, former Boston Police Commissioner Edward Davis testified that a "sergeant really has to have a strong basis of knowledge of all the rules and regulations and constitutional protections that are afforded the citizens of the Commonwealth to do the job properly," because when officers in the field "get confused and don't understand something, the first thing they do is call the sergeant." This "fundamental understanding" of "how things work," was a "critical component" of a sergeant's responsibilities, according to Commissioner Davis. And, the court supportably found, those skillsets were tested by the exam.

The Officers' reliance on *Beecher* is further undermined by the different approach taken in that case towards validation of the exam. We stated that for an exam to be valid, the court must be satisfied that "it demonstrably selects people who will perform better the required on-the-job behaviors after they have been hired and trained." *Id.* at 1021–22. We observed that "[t]he crucial fit is not between test and job lexicon, but between the test and job performance." *Id.* at 1022. This approach resembles what the Guidelines, adopted four years after *Beecher,* call "criterion-related validity." 29 C.F.R. § 1607.5(B) ("Evidence of the validity of a test or other selection procedure by a criterion-related validity study should consist of empirical data demonstrating that the selection procedure is predictive of or significantly

correlated with important elements of job performance."). Because in this case, as we have discussed, we assess validity for the most part under the separate "content validity" framework, *Beecher's* relevance is further limited.

None of the remaining arguments advanced by the Officers seriously support any claim that the exams are not materially better predictors of success than would be achieved by the random selection of ***117** those officers to be promoted to sergeant. The parties' arguments, instead, focus on how *much* better the exams were. Do they test enough skills and knowledge? Do they weigh the answers in an appropriate, valid way? In finding Outtz persuasive on these points, the district court as factfinder did not clearly err.[12]

B. Rank–Order Selection

When officials at the BPD received the results of the 2005 and 2008 sergeant promotional examinations from HRD, they selected as many police officers for promotion as there were vacancies currently available, beginning with the highest-scoring name at the top of the list and moving down the list, one at a time, in order of the score each candidate received. The Officers argue that this method of selection—quite independently from the written examination itself—led to a disparate impact and the district court was obligated to conduct a separate analysis of its validity under Title VII. We review the legal sufficiency of the district court's holding on this point de novo and its subsidiary fact-finding for clear error. *E.E.O.C. v. Steamship Clerks Union, Local 1066,* 48 F.3d 594, 603 (1st Cir.1995).

The Officers first argue that the district court failed altogether to wrestle with the consequences of rank-order selection. This is clearly not the case. Citing section 1607.14(C)(9) of the Guidelines, the district court noted in its exegesis of the law that "[t]he use of a ranking device requires a separate demonstration that there is a relationship between higher scores and better job performance." *Lopez,* 2014 U.S. Dist. LEXIS 124139, at *16–17. The court went on to find that Boston's selection method "reliably predicts a candidate's suitability for the job, such that persons who perform better under the test method are likely to perform better on the job." *Id.* at *61.

This finding by the district court, to the Officers, is "not enough." Based on their reading of the Guidelines, something more is required. The Officers argue that the use of the results of an examination that is "minimally valid" insofar as it tests job-related skills may not necessarily be valid if used to select candidates solely according to their scores on that exam.

Two provisions of the Guidelines discuss an employer's appropriate use of a rank-ordering selection method. In the section of the Guidelines establishing "General Principles," the EEOC has advised the following:

> The evidence of both the validity and utility of a selection procedure should support the method the user chooses for operational use of the procedure, if that method of use has a greater adverse impact than another method of use. Evidence which may be sufficient to support the use of a selection procedure on a pass/fail (screening) basis may be insufficient to support the use of the same procedure on a ranking basis under these guidelines. Thus, *if a user decides to use a selection procedure on a ranking basis, and that method of use has a greater adverse impact than use on an appropriate pass/fail basis* (see section 5H below), the user should have sufficient evidence of validity and utility to support the use on a ranking basis.

***118** 29 C.F.R. § 1607.5(G) (emphasis supplied). The Guidelines also contain a refinement of this principle specific to the use of content validity studies in the "Technical Standards" section:

> If a user can show, by a job analysis or otherwise, that a higher score on a content valid selection procedure is likely to result in better job performance, the results may be used to rank persons who score above minimum levels. Where a selection procedure supported solely or primarily by content validity is used to rank job candidates, the selection procedure should measure those aspects of performance which differentiate

among levels of job performance.

Id. § 1607.14(C)(9).

These two statements evidence some inconsistency. Section 1607.5(G) clearly indicates that an employer need have sufficient evidence of validity to support use of the exam on a ranking basis "*if* ... that method of use has a greater adverse impact than use on an appropriate pass/fail basis" (emphasis supplied). Under this guidance, if an exam is valid, one may use it on a rank-order basis unless the use of rank ordering creates or adds to a disparate impact. One can read section 1607.14(C)(9), however, as requiring that, to defend rank ordering, the employer must first show that "a higher score on a content valid selection procedure is likely to result in better job performance"; i.e., one must validate the use of ranking itself if the exam as a whole produces a disparate impact. Other provisions of the Guidelines support this latter reading, albeit without acknowledging the inconsistency. *Compare, e.g., id.* § 1607.5(G) ("[I]f a user decides to use a selection procedure on a ranking basis, *and* that method of use has a greater adverse impact than use on an appropriate pass/fail basis ..., the user should have sufficient evidence of validity and utility to support the use on a ranking basis." (emphasis supplied)), *with* Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, 44 Fed.Reg. 11,996, 12,005, Question and Answer n. 62 (1979) ("Use of a selection procedure on a ranking basis may be supported by content validity if there is evidence from job analysis or other empirical data that what is measured by the selection procedure is associated with differences in levels of job performance.").

Several courts have seemed to approach this issue by requiring more scrutiny of the validation evidence as a whole when rank ordering is used, particularly when the exams in question have led to closely bunched scores. *See Johnson v. City of Memphis,* 770 F.3d 464, 479–81 (6th Cir.2014), *cert. denied*, —— U.S. ——, 136 S.Ct. 81, 193 L.Ed.2d 34 (2015); *Police Officers for Equal Rights v. City of Columbus,* 916 F.2d 1092, 1102–03 (6th Cir.1990); *Guardians Ass'n of N.Y.C. Police Dep't, Inc. v. Civil Serv. Comm'n of City of N.Y.,* 630 F.2d 79, 100–05 (2d Cir.1980).

The district court in this case expressly adopted the approach most favorable to the Officers, citing 29 C.F.R. § 1607.14(C)(9), for the proposition that "[t]he use of a ranking device requires a separate demonstration that there is a relationship between higher scores and better job performance." *Lopez,* 2014 U.S. Dist. LEXIS 124139, at *16–17. As we have noted, *supra,* and as the Officers seem to ignore, the court then specifically found that it was "satisfied on the evidence that Boston carried its burden of showing" "that persons who perform better under the test method are likely to perform better on the job." *Id.* at *61–62. As a predicate to this finding, the district court observed that a group of incumbent sergeants who took an exam in 2005 that ***119** contained 53 of the questions asked of applicants on the sergeant's exam had a materially higher passing rate on those common questions than did the job applicants. *Id.* at *56–57. The district court viewed this evidence as showing that "those questions were related to the sergeants' actual performance of their jobs." *Id.* at *57. The Officers' only reply is to say that this evidence only shows that people who previously did well on the exam (and thus became sergeants) still did well on it. But the Officers point to no evidence that these incumbent sergeants in 2005 somehow achieved their positions by previously taking the same, or more or less the same, exam that was first offered in 2005.

Even accepting the district court's opinion that added scrutiny was called for because rank ordering was used, whatever added scrutiny one need apply here certainly falls short of the added scrutiny one would apply if rank ordering had been a material contributor to the disparate impact. Although they belatedly offer on appeal, without citation to the record, counsel's own calculations that "banding" in lieu of rank order selection would have caused more Black and Hispanic applicants to be "reachable" for selection by subjective "performance" criteria, the Officers made no effort to demonstrate that an increased number of Black and Hispanic applicants likely would have been selected under such an alternative approach. Rank ordering furthers the City's interest in eliminating patronage and intentional racism under the guise of subjective selection criteria. Such a goal is itself a reasonable enough business need so as to provide some weight against a challenge that is unaccompanied by any showing that rank order selection itself caused any disparate impact in this case.[13]

None of this is to suggest that Boston could not have come up with an exam that did a better job of furthering its goal of selecting the best candidates for promotion to the position of sergeant. The Officers argue persuasively that Boston could have made the exam more valid. Indeed, Outtz agreed and so, too, it would appear, does the City, which, counsel tells us, has since 2008 developed a new exam that it now uses.

The point, instead, is that the record contains detailed, professionally buttressed and elaborately explained

support for the district court's finding "that persons who perform better under the test method are likely to perform better on the job." *Id.* at *61. Given that plainly supported finding, it makes little sense to debate in the abstract how much better the exam might have been. Instead, it makes more sense to move to the next step of the inquiry to see if there is any alternative selection test that would have had less adverse impact. If so, then the court will have a meaningful gauge of validity by comparing the two tests. And if the alternative test with less adverse impact has equal or greater validity, it makes no difference how valid the employer's actual test is; the employee wins. *Ricci*, 557 U.S. at 578, 129 S.Ct. 2658 (citing 42 U.S.C. §§ 2000e–2(k)(1)(A)(ii) and (C)). Conversely, absent proof of an equally or more valid test that has less adverse impact, there is no reason to force the employer to promote randomly if the employer has a tool that will do meaningfully better than that. For this reason, once a court concludes that a selection device is materially more job-related than random selection would be, it makes ***120** sense to turn the focus sooner rather than later to the question of whether there is any alterative option that is as good or better, yet has less adverse impact. Otherwise, courts and litigants are left to engage in unpredictable qualitative assessments without any meaningful gauge as to what is enough. We therefore turn next to that question.

C. The Officers' Alternatives

So, the pivotal question on appellate review is whether the evidence compelled a finding "that the employer refuse[d] to adopt an available alternative employment practice that has less disparate impact and serves the employer's legitimate needs." *Id.* To carry this burden, plaintiffs must "demonstrate a viable alternative and give the employer an opportunity to adopt it." *Allen v. City of Chicago*, 351 F.3d 306, 313 (7th Cir.2003).

Outtz explained that he thought the Officers would be unlikely to carry this burden due to the very large number of applicants for relatively few open positions in Boston. On the 2008 exam, for example, where the disparate impact was much greater than in 2005, there were only 26 openings for 504 applicants. He explained that his experience is that:

> [I]n dealing with adverse impact[,] the ball game is played, for the most part, in terms of selection ratio. If I come to—if an employer comes to me and says, "Look, I've got five job openings and I've got 5,000 people that are applying for those five jobs and I want you to develop a system that reduces adverse impact," I'm just going home.

The Officers' own expert agreed that the selection ratio heavily influenced the menu of available options, offering his opinion that the degree of adverse impact caused by a selection process "depends so much on how many people you appoint."

The Officers baldly assert that the district court did not find "that Plaintiffs failed to meet their burden of putting forward a specific less discriminatory alternative." In fact, the district court precisely so found—twice. *Lopez*, 2014 U.S. Dist. LEXIS 124139, at *78 (holding that the Officers' showing was "not enough to carry their burden on this issue" and did not "demonstrat[e] by the evidence that there was an alternative employment practice with equal validity and less adverse impact that was available and that BPD refused to adopt").

The Officers also contend that "[i]t is undisputed that ... adding test components such as an assessment center, structured oral interview, or performance review to an exam process increases the validity of an exam while having less adverse impact on minorities." Yet the Officers failed to offer any evidence that would have compelled the district court to find that the deployment of any of these supposedly "undisputed" solutions would have led to "a smaller racial disparity in outcomes," *Jones*, 752 F.3d at 55, given the selection ratios facing authorities in Boston.

Our own review of the record does disclose testimony convincingly establishing that, as a general matter, incorporation of selection tools such as use of "hurdles," banding, oral interviews, so-called assessment centers, and open ended "situational judgment" questions generally tend to result in less adverse impact than does a reliance on multiple choice exams. What is missing, though, is any rebuttal to Outtz's opinion that the low rates of job openings in the Boston sergeant ranks relative to the number of applicants made it unlikely that any alternative selection device would have materially reduced adverse impact in 2005 and 2008.

***121** The Officers did offer evidence that the mean differentials on the oral portion of an exam Boston used in 2002 were less than the mean differentials on the written portions of that exam. But the 2002 exam as a whole still had substantially the same adverse impact as did the exams administered in 2005 and 2008.[14] And, again, the Officers provide no analysis of the effect of the selection ratios in 2005 and 2008.

Additionally, as the district court noted, Boston's prior attempt to employ assessment centers with situational

exercises and oral questioning in its 2002 promotional exam resulted in a cost of $1.2 million to develop the exam and the required "transporting, housing, and training a substantial number of police officers from throughout the country who acted as the assessors," *id.* at *70, without generating any convincing support that repeating such an approach in 2005 or 2008 would have reduced adverse impact, *id.* at *73. In concluding that the City was not required to again incur such costs without any demonstration that adverse impact would be materially reduced, the district court acted well within its discretion in making the judgments called for by the applicable law.[15] *See Watson,* 487 U.S. at 998, 108 S.Ct. 2777 (opinion of O'Connor, J.) ("Factors such as the cost or other burdens of proposed alternative selection devices are relevant in determining whether they would be equally as effective as the challenged practice in serving the employer's legitimate business goals.").

Satisfying a plaintiff's burden on this point at trial "demands evidence that plaintiffs' preferred alternative would have improved upon the challenged practice," *Johnson,* 770 F.3d at 477, not just that such practices exist in the abstract. Furthermore, securing the reversal of a trial court's factual finding that the Officers' proof on this point was not persuasive required evidence that is so compelling as to render its rejection clear error. The Officers' scattershot listing of alternatives without any developed rejoinder to Outtz's testimony concerning the challenge posed by the selection ratios in 2005 and 2008 fell short of this mark.[16]

III. Conclusion

Given our finding that the district court applied the correct law and committed no ***122** clear error in finding persuasive the expert evidence tendered by Boston, we *affirm* the district court's order finding that the exams Boston used in 2005 and 2008 did not violate Title VII and we therefore *affirm* as well the entry of judgment in favor of all defendants.

TORRUELLA, Circuit Judge, concurring in part and dissenting in part.

I agree with my colleagues in the majority only to the extent that the challenged tests did have a disparate impact. There is little doubt in my mind, however, that the majority's question, whether "the employer[s] show[ed] that the challenged employment practice creating this disparate result is nevertheless job-related for the position in question and consistent with business necessity," *supra* at 111, cannot be answered in the affirmative based on this record.[17] To my view, the district court committed clear error in finding that the challenged tests were valid when placed under the legal prism of Title VII, 42 U.S.C. § 2000e *et seq. M.O.C.H.A. Soc'y, Inc. v. City of Buffalo,* 689 F.3d 263, 275 (2d Cir.2012); *Ass'n of Mex.–Am. Educators v. California,* 231 F.3d 572, 584–85 (9th Cir.2000) (en banc); *Melendez v. Ill. Bell Tel. Co.,* 79 F.3d 661, 669 (7th Cir.1996); *Hamer v. City of Atlanta,* 872 F.2d 1521, 1526 (11th Cir.1989); *Bernard v. Gulf Oil Corp.,* 890 F.2d 735, 739 (5th Cir.1989).

A review of the record shows that Boston[18] did not, contrary to the district court's finding and the majority's assertion, "show[ ] 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." *Supra* at 111 (quoting 29 C.F.R. § 1607.5(B)); *see also* 29 C.F.R. § 1607.5(A). Because there is ample precedent on which to draw, *see, e.g., Bos. Chapter, NAACP, Inc. v. Beecher,* 504 F.2d 1017 (1st Cir.1974), I need not engage the majority's emphasis on the non-binding nature of EEOC Guidelines, *supra* at 111–12, nor rest my objection on what I would consider the Guidelines' rather overwhelming persuasiveness vis-à-vis this case. *Id.* at 111–12 (citing *Jones v. City of Bos.,* 752 F.3d 38, 50 n. 14 (1st Cir.2014)). It is enough to say that, based on our precedent and this record, there is a solid legal basis to find that the district court's acceptance of Boston's case for content validity is clearly erroneous.

The most significant flaws in Boston's case for validity should each independently have been fatal to it: Boston failed to demonstrate (1) that the 1991 Validation Report and 2000 job analysis were applicable and reliable[19] and (2) that the exams tested "representative" and critical knowledge, skills, and abilities ("KSAs") necessary to quality for the position of police sergeant.

This first flaw stems from "the way in which the validation study was performed" and its effect on test validity. *Beecher,* 504 F.2d at 1025. The Validation Report and job analysis were defective. The district court acknowledged the "rule of ***123** thumb" that a job analysis should typically have been performed within the last five to eight years to be reliable. *López v. City of Lawrence,* No. 07–11693–GAO, 2014 U.S. Dist. LEXIS 124139, at *51 (D.Mass. Sept. 5, 2014). Yet, the 1991 job analysis and resultant Validation Report predate the first of the contested exams by fourteen years. Neither of the two conditions noted by the district court as potentially saving an older analysis from obsolescence—lack of change in job requirements or a later review updating the

analysis—rescue the Report. *Id.; cf.* 29 C.F.R. § 1607.5(K) (explaining totality of circumstances should be considered in determining whether a validation study is outdated).

The Officers bolstered the presumption that a test more than eight years old is not reliable, and the common sense conclusion that a position changes over time, by pointing to specific evidence that defendants' police departments changed practices since the Report and analysis were performed: testimony from Commissioner Edward F. Davis that Lowell implemented a community policing model and a 2002 Boston Commissioner's memo referring to changes in policing policy and practice. While the district court was entitled to rely on Dr. Outtz's testimony as to the unchanging nature of the position of sergeant, it clearly erred in doing so for the proposition it drew from his testimony, that the position of police sergeant in the *defendant* departments had not changed, as Dr. Outtz based his statement on "[his] experience generally" regarding the position in other municipalities, including those in other states.

The subsequent job analysis completed in 2000, within the time range to be presumed reliable, is unreliable by virtue of the way it was performed. The 2000 job analysis suggests that the eleven subject matter experts ("SMEs"), sergeants and detective sergeants, relied upon by the testing firm to evaluate KSAs and tasks for inclusion in the exam, were to do so individually; the analysis details procedures for reconciling disparate results to determine which items should make the cut. For example, "[f]or a KSA to be included as a [sic] important component of the Police Sergeant position, the KSA had to be rated by nine ... of the eleven ... SMEs" in a certain way across all five categories. Yet the eleven SMEs evaluating 160 KSAs each rated all 160 KSAs' five attributes—job relatedness, time for learning, length of learning, differential value to performance, and necessity[20]—in *exactly the same way*, although there were 72 possible ways to rate *each* KSA. The same was true of task ratings, wherein each SME was supposed to rate each of 218 tasks' frequency, importance, necessity, relationship to performance, and dimensions,[21] despite the fact that each of 218 tasks could be rated in 1,674 ways. I will not speculate as to how and why this total agreement occurred but only observe that ***124** an analysis that generates a result so unfathomably inconsistent with its proposed methods is not reliable.[22] As such, it was clear error to find the 2000 job analysis supports the exams' validity. *Beecher*, 504 F.2d at 1025.

Beyond these threshold issues, the resultant exams did not test a representative portion of KSAs. *See* 29 C.F.R. § 1607.5(B). Nor did they test critically important KSAs "in proportion to their relative importance on the job." *Guardians Ass'n of N.Y.C. Police Dep't, Inc. v. Civil Serv. Comm'n of N.Y.C.,* 633 F.2d 232, 243–44 (2d Cir.1980) (citation omitted); *see also Beecher*, 504 F.2d at 1024 (noting district court did not err in finding that two significant correlations between exam and job performance components did not make " 'convincing' evidence of job relatedness" (citation omitted)); *see also* 29 C.F.R. § 1607.14(C)(2) (an exam should measure "critical work behavior(s) and/or important work behavior(s) constituting most of the job").

The 2000 job analysis identified 163 "important tasks" and 155 "important" KSAs. The district court acknowledged that the eighty-point multiple-choice portion of the exams tested primarily the "K" of the KSAs, knowledge, and failed to measure key skills and abilities, and thus would not be independently valid. *López*, 2014 U.S. Dist. LEXIS 124139, at *60–61. The E & E component that purportedly compensated for the "SA" deficit, edging the exams into the realm of validity, consisted of a single sheet requiring candidates to bubble in responses as to length of work experience in departmental positions by rank, educational background, and teaching experience. As the majority concedes, this component had a minimal effect on score. *Supra* at 113.

The conclusion that more than half, *López*, 2014 U.S. Dist. LEXIS 124139, at *54, or nearly half, *supra* at 115 n. 11, of applicable KSAs were or could be tested by the exams overestimates the number of KSAs tested by the E & E component. But even if that estimate were correct, relying upon this quantitative measure misses that representativeness is partly qualitative.

It is quite a stretch to conclude that the E & E's bubbles incorporated measures of the majority of key skills and abilities. It is even more difficult to conclude from the record that the skills and abilities measured received representative weight. *Supra* at 113. How, exactly, could this worksheet test, as the testability analysis suggests, "[k]nowledge of the various communities within the Department's jurisdiction and the factors which make them unique," "[s]kill in perceiving and reacting to the needs of others," or "[k]nowledge of the procedures/techniques when a major disaster occurs,"? And how, if it only affected the ultimate score by five to seven percent at most, *supra* at 113, could it be said that the KSAs for which the E & E ostensibly tested were adequately represented relative to those KSAs tested on the multiple-choice component?

The exam's failure to include particularly significant KSAs also precludes representativeness. *See Gillespie v. Wisconsin*, 771 F.2d 1035, 1044 (7th Cir.1985) ("To be representative for Title VII purposes, an employment test

must neither: (1) focus exclusively on a minor aspect of the position; ***125** nor (2) *fail to test a significant skill required by the position.*" (emphasis added)); *Guardians*, 630 F.2d at 99. The exams here may have tested the knowledge a supervisor must have but omitted any meaningful test of supervisory skill, which is unquestionably essential to the position of police sergeant. *López*, 2014 U.S. Dist. LEXIS 124139, at *51. Written tests of supervisory skill have been found by other courts to be altogether inadequate to evaluate that attribute. *See Vulcan Pioneers, Inc. v. N.J. Dep't of Civil Serv.,* 625 F.Supp. 527, 547 (D.N.J.1985), *aff'd on other grounds,* 832 F.2d 811, 815–16 (3d Cir.1987); *see also Firefighters Inst. for Racial Equal. v. City of St. Louis*, 549 F.2d 506, 513 (8th Cir.1977).

As in *Beecher*, "[t]here are, in sum, too many problems with the test ... to approve it here." 504 F.2d at 1026. It cannot be anything but clear error, *supra* at 114, to find valid exams based on an outdated validation report and a facially flawed job analysis, exams that are not only unrepresentative but also omit critical KSAs for the position of police sergeant. To endorse the means by which these exams were created and the exams themselves here establishes a perilous precedent that all but encourages corner-cutting when it comes to Title VII.

On these grounds, I respectfully dissent.

All Citations

823 F.3d 102

Footnotes

[1] This agency is the Human Resources Division of the Massachusetts Executive Office of Administration and Finance. *Lopez v. City of Lawrence,* No. 07–11693–GAO, 2014 U.S. Dist. LEXIS 124139, at *7 n. 1 (D.Mass. Sept. 5, 2014).

[2] The district court offered a detailed summary of this litigious history. *See Lopez,* 2014 U.S. Dist. LEXIS 124139, at *24–27.

[3] The Officers argue that Boston misrepresented its reliance on the 1991 report and that the City, in fact, used only a less-thorough report conducted in 2000. The Officers' evidence for this consists of a comparison, in a footnote in their appellate brief, between three tested skill areas out of fifteen total areas on the 2008 outline of exam questions and similar language from the 2000 job analysis. We decline to find that this perfunctory, post-judgment sampling demonstrates that the district court committed clear error.

[4] Boston supplemented the HRD-produced examination with additional jurisdiction-specific questions that sought to probe a candidate's knowledge of Boston-specific rules, orders, and regulations.

[5] The Officers point out that the same E & E sheet was used to identify candidates for promotion among Massachusetts firefighters in 2010.

[6] State law permitted a certain amount of flexibility for municipalities to "bypass" a candidate who had the next-highest score on the ranked list. Mass. Gen. Laws ch. 31, § 27. The municipality could be held accountable to the bypassed employee and, if challenged, would have to articulate a defensible reason for skipping him or her over. *See City of Cambridge v. Civil Serv. Comm'n,* 43 Mass.App.Ct. 300, 682 N.E.2d 923, 925 (1997). No justification "inconsistent with basic merit principles, can[ ] be used to justify a bypass," including a candidate's race. *Mass. Ass'n*

of Minority Law Enf't Officers v. Abban, 434 Mass. 256, 748 N.E.2d 455, 462 (2001). The Massachusetts Bay Transit Authority ("MBTA"), a state agency and a defendant, behaved slightly differently during the relevant years by treating all the candidates on HRD's list as having scored equally and narrowing down their pool of candidates by using oral interviews.

[7] The other defendants did not concede that the statistical analyses applied to the outcomes among their smaller groups of applicants established a disparate impact, and the district court agreed with the defendants. Our disposition of this appeal does not require us to assess the correctness of that ruling.

[8] The district court was entitled to rely on this conclusion, despite the Officers' various quibbles with the methodologies used to compile the 1991 and 2000 reports.

[9] The Officers place this variation slightly lower, at 1% to 4%, relying on testimony suggesting that no candidate could reach the ceiling of the potential boost offered by the E & E. Unguided by fact-finding on this narrow question, we note only the absence of any evidence that Outtz's opinion turned on a plainly erroneous calculation of the precise percentage.

[10] Joined by the United States as amicus curiae, the Officers further dispute the "linkage" between these validation reports—both the 1991 and 2000 reports—and the examinations themselves. Their chief challenge on this front revolves around a "testability analysis" document prepared in connection with the 1991 report that evaluates which key skills could, in theory, be tested on a future examination but does not directly link the skills identified to actual examination or E & E content. The defect with the Officers' reliance on this document is that it asks us to pretend that it was the only relevant evidence the district court could rely on in drawing a connection between the validation reports and the examinations as administered. This was hardly the case. The district court weighed the testimony of Dr. Wiesen and Dr. Outtz, both of whom had analyzed the examinations as well as the reports, reviewed the testability analysis, applied their scientific expertise, and formed their own (differing) judgments as to whether the examinations tested the skills identified by the reports. In crediting Outtz's testimony, the district court did not clearly err.

[11] In the district court's observation that "more than half of the KSAs identified as pertinent to the job were tested," *Lopez,* 2014 U.S. Dist. LEXIS 124139, at *54, the Officers see clear error, pointing out that the 1991 testability analysis only identified 70 out of a total 156 critical KSAs (i.e., not quite half) that could be tested on the exam. We decline the Officers' invitation to find this difference to be so material as to constitute clear error.

[12] The Officers did not move to strike any portion of Outtz's testimony under *Daubert v. Merrell Dow Pharmaceuticals, Inc.,* 509 U.S. 579, 113 S.Ct. 2786, 125 L.Ed.2d 469 (1993). Hence, even if we had thought that any part of Outtz's opinion was unreliable or unsupported, we would have had to employ plain error review. *See United States v. Diaz,* 300 F.3d 66, 74 (1st Cir.2002).

[13] Given the absence of any showing that an equally or more valid alternative to rank-order selection would have

reduced disparate impact, we need not address the Officers' arguments that any state law favoring rank order selection is unlawful or preempted.

[14] The adverse promotional impact ratio in 2002 was calculated to be .32. In 2005, it was .28.

[15] Boston had previously tried other tactics to reduce adverse impact. In 1992 and 2002 Boston experimented by integrating an assessment center component into the exam. After the 1992 exam, the City used its bypass authority to promote several Black candidates over Caucasian candidates in order to achieve compliance with a consent decree and the Guidelines. They were sued and the bypasses were reversed. *See Abban,* 748 N.E.2d 455.

[16] The Officers' failure to explain how a particular alternative would have reduced disparate impact in 2005 and 2008—and by how much—is particularly odd given the obvious mootness of their claim for injunctive relief. Consequently, had the remedy phase of trial proceeded as the Officers would have hoped, each officer would have needed to show that, more likely than not, he or she would have been promoted had Boston used an equally or more valid selection tool with less impact. *See* 42 U.S.C. § 2000e–5(g)(1) (authorizing "back pay" remedy for Title VII violation); *Azimi v. Jordan's Meats, Inc.,* 456 F.3d 228, 235 (1st Cir.2006) ("Injuries allegedly caused by the violation of Title VII ... must be proven to the factfinder ... which may reasonably find, within the law, that while there has been [injury], the plaintiff has not been injured in any compensable way by it."). How any officer could have made such a showing without first securing a liability finding predicated on a specific alternative selection tool that would have been equally or more valid and produced less adverse impact is entirely unclear.

[17] I would also have found the Officers established a prima facie case as to all defendants, but, as the majority does not address this question, *supra* at 107, I will focus on test validity.

[18] Like the majority, *supra* at 107, I will refer primarily to Boston for the sake of simplicity.

[19] As I would find neither sufficed to support the exams' validity, it does not matter which Boston relied upon for each test, the 2000 job analysis or 1991 Validation Report. *See supra* at 109 n. 3.

[20] Job relatedness could be answered "[y]es" or "[n]o"; time for learning, "[b]efore assignment" or "[a]fter assignment"; length of learning, "[l]onger than brief orientation" or "[b]rief orientation"; differential value to performance, "[h]igh," "[m]oderate," or "[l]ow"; and necessity, "[r]equired," "[d]esirable," or "[n]ot required."

[21] Frequency could be rated "[r]egular[ ]," "[p]eriodic[ ]," or "[o]ccasional[ ]"; importance, "[v]ery important," "[i]mportant," or "[n]ot important"; necessity, "[n]ecessary upon entry" or "[n]ot necessary"; and relationship to performance, "this task clearly separates the best workers," "better workers seem to perform this better than poor or marginal workers," or "[m]ost perform this task equally well." Dimensions could be answered using any combination of "[o]ral [c]ommunication," "[i]nterpersonal [s]kills," "[p]roblem ID & [a]nalysis," "[j]udgment," and

"[p]lanning and [o]rganizing" or "all."

[22] A second suspect aspect of this analysis, one that further clarifies how troubling the purported across-the-board agreement is, is in *how* the SMEs rated certain KSAs and tasks. For example, all eleven SMEs—including two assigned to administrative roles,—responded that "[s]et [ting] up command posts at scenes of[ ]robberies, homicides, fires, etc.," was a "daily" task.

End of Document