

File Name: 14a0271p.06

UNITED STATES COURT OF APPEALS

FOR THE SIXTH CIRCUIT

MARILYN JOHNSON, et al.,

Plaintiffs-Appellants/Cross-Appellees,

v.

CITY OF MEMPHIS,

Defendant-Appellee/Cross-Appellant.

Nos. 13-5452/5454

Appeal from the United States District Court
for the Western District of Tennessee at Memphis.
Nos. 2:00-cv-02608; 2:04-cv-02013; 2:04-cv-02017—S. Thomas Anderson, District Judge.

Argued: January 30, 2014

Decided and Filed: October 27, 2014

Before: SUHRHEINRICH, GIBBONS, and COOK, Circuit Judges

COUNSEL

ARGUED: David M. Sullivan, Memphis, Tennessee, for Appellants/Cross-Appellees. Louis P. Britt, FORD & HARRISON LLP, Memphis, Tennessee, for Appellee/Cross-Appellant. **ON BRIEF:** David M. Sullivan, Memphis, Tennessee, for Appellants/Cross-Appellees. Louis P. Britt, J. Dylan King, Joshua J. Sudbury, FORD & HARRISON LLP, Memphis, Tennessee, for Appellee/Cross-Appellant.

OPINION

COOK, Circuit Judge. After more than thirteen years of litigation, including a bench trial, numerous preliminary injunctions, and a previous appeal affirming the grant of injunctive relief for some plaintiffs, *see Johnson v. City of Memphis* (“*Johnson Appeal I*”), 444 F. App’x

856, 861 (6th Cir. 2011), three consolidated cases challenging the City of Memphis’s (“City”) police promotional processes as racially discriminatory return on cross-appeals. The appeals address two allegedly discriminatory sergeant promotional processes that occurred in 2000 and 2002 (the “2000 process” and “2002 process”¹), targeting three matters decided by the district court at different phases of the litigation: (1) the order dismissing plaintiffs’ negligence claim concerning the already-invalidated 2000 process under Tennessee’s governmental-immunity statute, Tenn. Code Ann. § 29-20-205; (2) the bench-trial decision invalidating the 2002 process for violating Title VII’s disparate-impact prohibition, *see* 42 U.S.C. § 2000e-2(k)(1); and (3) the final judgment and related orders awarding back pay and interest to plaintiffs and more than \$1 million in fees and expenses to their attorneys. Both the plaintiffs and the City appeal various aspects of these decisions.

For the following reasons, we affirm in part and reverse in part the district court’s judgment, and we remand the fees issues for further consideration.

I. BACKGROUND

We briefly summarize the factual background of these cases thoroughly detailed in the district court’s bench-trial opinion. The City’s promotional processes have engendered controversy for nearly forty years, prompting numerous lawsuits alleging racial and gender discrimination by such parties as the United States Department of Justice, the Afro-American Police Association, and white and minority officers. *See Aiken v. City of Memphis*, 37 F.3d 1155, 1158–60 (6th Cir. 1994) (en banc) (detailing the extensive litigation history). Despite the City’s repeated assurances of adopting race-neutral promotional processes, we observed that, as of the mid-1990s, “incredibly, the City continue[d] to make police and fire department promotions according to procedures that have not been validated as racially neutral.” *Id.* at 1164.

The City responded with a 1996 promotional process (“1996 process”) designed by Dr. Mark Jones, an industrial and organizational psychologist, and overseen by a Department of Justice consultant. The 1996 process consisted of four components, weighted as follows: a “high-fidelity” law enforcement role-play exercise, 50%; written test, 20%; performance

¹We refer to the second promotion period as the “2002 process,” even though the City administered the test in September 2001, for consistency with the parties’ arguments and our previous decision.

evaluations, 20%; and seniority, 10%. Arbitration proceedings involving claims under the City's Memorandum of Understanding with the police union ensued, but no Title VII litigation resulted.

Dr. Jones modeled the City's next promotion protocol after the 1996 process, replacing the role-play component with a video-based practical test because of security and practicability concerns. The 1996 simulation had taken more than two months (testing and scoring) to evaluate individually more than 400 candidates, and the City discovered problems with candidate coaching during the exercise. The following components initially comprised the 2000 process: a "low-fidelity" (*i.e.*, no role-play) video-based practical test, 50%; job knowledge test, 20%; performance evaluations, 20%; seniority, 10%. After the City discovered that leaked answers compromised the results of the video test, the City excluded the video test and reweighted the remaining test components. The adjustments to the 2000 process prompted the first of these disparate-impact cases, *Johnson v. City of Memphis*, No. 00-2608, and the City ultimately consented to the invalidation of the 2000 process by Judge Jon McCalla in June 2001. (*See* R. 58, Order at 1-2.²)

Attempting to avoid the test-security issues encountered in the previous two promotional periods, the City hired outside consultants Jeanneret & Associates to design the replacement tests that would become the 2002 process. After the City submitted a testing proposal to the district court, Judge McCalla held a status conference to hear plaintiffs' objections and instructed plaintiffs' expert to work with the City's expert, Dr. Richard Jeanneret. The City addressed the concerns raised by plaintiffs' expert, and the district court granted the City's motion to proceed with the 2002 process. The 2002 process included the following equally weighted test components: an investigative logic test; a job-knowledge test; an application-of-knowledge test; a grammar and clarity test; and a "low-fidelity" video-based practical test.

The City administered the 2002 process to 517 applicants between September 27-29, 2001, and completed grading in fall 2002. Raw scores ranged from 174.75-358.75 out of a possible 384.5 points. The City converted these scores to a 100-point scale and then—honoring an agreement with the officers' union—added up to 10 points for seniority to the final promotion score. Promotion scores ranged from 53.511-103.303, of a possible 110 points. Despite the

²All record citations refer to case No. 00-2608.

City's efforts, the 2002 process resulted in minority candidates scoring disproportionately worse than white candidates. Using Dr. Jeanneret's rank-ordered promotion scores, the City promoted 86 of the 274 African-American candidates (31.4%) and 176 of the 240 white candidates (73.3%). The original plaintiffs amended their pleadings to challenge the disparate impact of the 2002 process, and two additional lawsuits—*Johnson v. City of Memphis*, No. 04–2017, and *Billingsley v. City of Memphis*, No. 04–2013—joined the consolidated proceedings, which had been reassigned to then-District Judge Bernice Donald in September 2001.

The district court held a bench trial in July 2005 and issued its decision in December 2006. Its Memorandum Opinion and Order on Remedies rejected all claims except plaintiffs' Title VII disparate-impact claims as to the 2002 process. The court found that, while the 2002 sergeant test was valid and reliable, less discriminatory valid alternatives were available and, thus, the 2002 process violated Title VII. Though the court ordered the promotion of all minority plaintiffs, with back pay and seniority, it denied plaintiffs' request, at that time, to compete for promotion to the rank of lieutenant because they lacked the requisite two years' experience as sergeant. *See Johnson Appeal I*, 444 F. App'x at 857 (detailing district court's procedural history).

Following the bench-trial decision, the district court fielded a variety of remedies-related motions for injunctions and stays between 2007 and 2010. Because so much time had passed since the problematic 2000 and 2002 processes, plaintiffs' alleged injuries, in terms of lost pay and seniority, spilled over into subsequent promotional processes, as plaintiffs were denied the opportunity to apply for additional promotions. At different points, court orders relying on the Title VII judgment invalidating the 2002 process permitted plaintiffs to participate in those promotions, *see generally Johnson Appeal I*, 444 F. App'x at 857 (lieutenant promotions), but the district court repeatedly denied plaintiffs' request for additional retroactive seniority and back pay.

In March 2010, the court entered a preliminary injunction ordering the immediate promotion to the rank of lieutenant of 28 plaintiffs with passing exam scores and sufficient work experience, and we affirmed in *Johnson Appeal I*, 444 F. App'x at 857–58, 861. In affirming the preliminary injunction, the panel expressed “concern[] at the degree of delay” of “this case, now

in its eleventh year,” and admonished that it would entertain a mandamus petition if the district court failed to enter a final judgment within the next six months. *Id.* at 861 (noting that the district court’s 2006 bench-trial decision “remains interlocutory almost five years later”). After plaintiffs petitioned for mandamus in January 2013, the district court awarded back pay, interest, and attorneys’ fees and entered a final judgment, whereupon plaintiffs voluntarily dismissed their mandamus action.

The plaintiffs appeal the immunity-based denial of their negligence claim related to the 2000 process and various remedies and attorneys’ fees issues related to the 2000 and 2002 processes; the City cross-appeals the district court’s Title VII judgment invalidating the 2002 process and the related million-dollar attorneys’ fees award; and the plaintiffs present an alternative legal justification³ for the Title VII judgment against the 2002 process.

II. *JOHNSON I* PLAINTIFFS’ APPEAL: NEGLIGENCE CLAIM, 2000 PROCESS

First, the non-minority *Johnson I* plaintiffs dispute the application of governmental immunity to their negligence claim, targeting the already-invalidated 2000 process. They press this claim—their only one seeking damages—arguing that the decisionmakers responsible for the 2000 process committed non-discretionary acts ineligible for immunity. We review the district court’s grant of summary judgment de novo. *Ciminillo v. Streicher*, 434 F.3d 461, 464 (6th Cir. 2006).

According to the *Johnson I* plaintiffs, City officials violated a key provision of the City Charter requiring the use of “practical tests” in the promotion process. Specifically, they object to the City’s exclusion of the interactive, video-based component of the 2000 process upon discovering that some candidates received advance notice of the questions.

The district court rejected this argument, finding that “the decisions concerning what type of test to use, how to weight the various testing components, and how the tests are to be

³Though styled a “conditional cross-appeal” in plaintiffs’ response brief, we construe the argument as an alternative legal justification for the district court’s judgment. *See ASARCO, Inc. v. Sec’y of Labor*, 206 F.3d 720, 722 (6th Cir. 2000) (“It is a well settled principle that a prevailing party cannot appeal an unfavorable aspect of a decision in its favor.”); *see also Freeze v. City of Decherd*, 753 F.3d 661, 664 (6th Cir. 2014) (“Appellate courts reviewing grants of summary judgment may affirm on any grounds supported by the record.”); *Abel v. Dubberly*, 210 F.3d 1334, 1338 (11th Cir. 2000) (applying similar standard to post-trial motions for judgment as a matter of law, considering preserved alternative legal arguments).

administered are left to the discretion of the director of personnel,” and noting that the Charter’s practical-test requirement “must be interpreted by those in a position to make such decisions for [the City].” We agree with the district court.

Tennessee’s Governmental Tort Liability Act (GTLA) immunizes the state’s public officials from negligence suits where “the injury arises out of . . . [t]he exercise or performance . . . of a discretionary function, whether or not the discretion is abused.” Tenn. Code Ann. § 29-20-205(1). Tennessee courts measure the scope of this immunity with the “planning-operational test.” *Giggers v. Memphis Hous. Auth.*, 363 S.W.3d 500, 507 (Tenn. 2012). Because arguably “every act involves discretion,” courts must “examin[e] (1) the decision-making process and (2) the propriety of judicial review of the resulting decision.” *Bowers v. City of Chattanooga*, 826 S.W.2d 427, 431 (Tenn. 1992). Whereas discretionary “planning decision[s] usually involve[] consideration and debate regarding a particular course of action by those charged with formulating plans or policies,” non-discretionary “[o]perational decisions . . . implement preexisting laws, regulations, policies, or standards” and “do[] not involve the formulation of new policy.” *Giggers*, 363 S.W.3d at 507–08. Accordingly, we must determine whether the City Charter and ordinance prescribe sufficient instructions such that the formulation and modification of the 2000 process can be deemed operational, as opposed to discretionary.

Contrary to the *Johnson I* plaintiffs’ suggestion, the City Charter and related ordinance do not require “practical tests.” Rather, they provide that employment examinations “shall be of a *practical nature* and relate to such matters as will fairly test the relative competency of the applicant to discharge the duties of the particular position.” (R. 656-25, City Charter § 250.1 (emphasis added); *accord* R. 656-26, Civil Service Ordinance § 9-3.) This subtle difference suggests that the regulations provide a broad instruction that examinations test actual job functions, instead of a strict requirement for a specific type of interactive exercise, like a simulation or video-based test. Other aspects of the Charter provision similarly support treating test-design as a discretionary function. (*See* R. 656-25, City Charter § 250.1 (requiring “competitive job-related examinations under such rules and regulations as may be adopted by the Director of Personnel,” and providing that the exams “should be developed in conjunction with other tools of personnel assessment and . . . sound programs of job design to aid significantly in

the development and maintenance of an efficient work force and in the utilization and conservation of human resources”).) Plaintiffs offer no authority supporting their narrow interpretation. Nor do they explain how the Charter and ordinance preclude the City from taking the sensible step of voiding a compromised component of its employment examination.

The district court correctly recognized that City officials must interpret and implement the Charter’s broad guidance in devising fair and effective promotional processes. In the absence of specific regulations confining the City’s discretion, GTLA immunity shields this discretionary decision. *See Giggers*, 363 S.W.3d at 507–08. We therefore AFFIRM the district court’s grant of partial summary judgment to the City on this claim.

III. CITY’S CROSS-APPEAL: TITLE VII JUDGMENT, 2002 PROCESS

Next, the City cross-appeals the district court’s bench-trial ruling finding a Title VII disparate-impact violation. The parties agree that plaintiffs presented a prima facie case of the 2002 process’s disparate impact; the City promoted 264 of the 517 candidates, with a substantial disparity between the success rate of non-minority (175/240) and African-American candidates (86/274). The City argues, however, that the court applied an unduly deferential legal standard in finding that plaintiffs showed less discriminatory alternatives to the 2002 process. We review the court’s legal conclusions de novo and findings of fact for clear error. *E.g., Beaven v. U.S. Dep’t of Justice*, 622 F.3d 540, 547 (6th Cir. 2010).

A. *The Title VII Disparate-Impact Standard*

Though Title VII disparate-impact claims originated with the Supreme Court’s decision in *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971), Congress codified the disparate-impact standard in the Civil Rights Act of 1991. *See* 42 U.S.C. § 2000e-2(k)(1); *Ricci v. DeStefano*, 557 U.S. 557, 577–78 (2009). Courts assess the viability of these claims using a three-step burden-shifting framework akin to the familiar *McDonnell-Douglas* standard. *See* 42 U.S.C. § 2000e-2(k)(1)(A)–(k)(1)(C); *Black Law Enforcement Officers Ass’n v. City of Akron*, 824 F.2d 475, 480 (6th Cir. 1987).

[First,] a plaintiff establishes a prima facie violation by showing that an employer uses “a particular employment practice that causes a disparate impact on the basis of race, color, religion, sex, or national origin.” 42 U.S.C. § 2000e-2(k)(1)(A)(i).

[Second, the] employer may defend against liability by demonstrating that the practice is “job related for the position in question and consistent with business necessity.” *Ibid.* [Third,] . . . if the employer meets that burden, . . . [the] plaintiff may still succeed by showing that the employer refuses to adopt an available alternative employment practice that has less disparate impact and serves the employer’s legitimate needs. §§ 2000e-2(k)(1)(A)(ii) and (C).

Ricci, 557 U.S. at 578; *see also Davis v. Cintas Corp.*, 717 F.3d 476, 494–95 (6th Cir. 2013).

The City contests plaintiffs’ step-three showing of less discriminatory alternatives. To satisfy this element, the plaintiff must demonstrate: (1) the availability of alternative procedures that serve the employer’s legitimate interests and (2) produce “substantially equally valid” results, but with (3) less discriminatory outcomes. 29 C.F.R. § 1607.3(B); *see also Watson v. Fort Worth Bank & Trust*, 487 U.S. 977, 998 (1988); *Shollenbarger v. Planes Moving & Storage*, 297 F. App’x 483, 486–87 (6th Cir. 2008). As with Title VII claims of intentional discrimination, disparate-impact plaintiffs bear the burdens of production and persuasion at this step. 42 U.S.C. §§ 2000e(m), 2000e-2(k)(1)(A)(i)–(ii). Consequently, plaintiffs may not rest on speculation regarding the availability, validity, or less discriminatory nature of their proffered alternatives. *See, e.g., Allen v. City of Chicago*, 351 F.3d 306, 313, 316–17 (7th Cir. 2003) (deeming insufficient “vague or fluctuating” alternatives, and finding that the plaintiffs failed to substantiate their “bare assertion” of valid, less discriminatory alternatives); *Shollenbarger*, 297 F. App’x at 487 (emphasizing that “[t]he plaintiffs [a]re obligated to prove equally effective alternatives,” and that “[t]he purpose of [step three] is not to second guess the employer’s business decisions”).

B. Components of the 2002 Process & Plaintiffs’ Proposed Alternatives

As noted above, the 2002 process consisted of five testing components: (1) a “low-fidelity” video test, which required oral responses to video depictions of law enforcement scenarios; (2) an investigative logic test, consisting of multiple-choice and short-answer questions; (3) an open-book job-knowledge test; (4) an application test, with weighted scores differentiating between the most and least effective responses; and (5) a written communications exam testing for grammar and clarity.

As they did before the district court, plaintiffs assert three available alternatives to improve the 2002 process: (1) the 1996 process's high-fidelity role-playing exercise, which required candidates to respond to simulated law-enforcement scenarios ("1996 simulation"); (2) assessments of candidates' "integrity" and "conscientiousness"; and (3) a merit-promotion system similar to one used by the Chicago Police Department, which consists of interviews by merit-review boards. Yet, in arguing before this court for these alternatives, they shirk their duty to demonstrate the benefits of the Chicago-plan and integrity/conscientiousness theories, defending only the 1996 simulation as equally valid and less discriminatory. (Third Br. at 31–37.) Similarly problematic, plaintiffs neglect to explain how any of these alternatives would fit into the 2002 process, but we gather that they would either replace or complement its existing components.

Plaintiffs vouch for the 1996 simulation by pointing to its past success, including a sterling validation report documenting its non-discriminatory results. They also tout its benefits compared to the less practical (i.e., less like actual job duties), low-fidelity video test used in the 2002 process. Finally, they rely on their expert's claim that the 1996 simulation is more valid than the 2002 tests and "easily replicated." (*See* Third Br. at 32–35; R. 648-13, Trial Tr. (DeShon) at 1681–82; *see also* R. 648-15, Trial Tr. (DeShon) at 1848 (likening the difference between high-fidelity simulations and low-fidelity response exercises to "knowing versus doing").)

C. The District Court's Bench-Trial Findings Regarding Available Alternatives

After summarizing the proffered alternatives, which the court characterized as "broad suggestions [of] alternative testing modalities," the court found that plaintiffs satisfied the step-three burden of demonstrating available, equally valid, less discriminatory alternatives. It reasoned as follows:

It is of considerable significance that the City had achieved a successful promotional program in 1996 and yet failed to build upon that success. While the 1996 process was not perfect it appears to have satisfied all of the legal requirements of promotional processes. The 2000 process departed substantially from the 1996 model in its abandonment of the practical exercise and re-weighting of the remaining elements. The 2002 processes, while arguably more

sophisticated than its predecessors, suffered from a grossly disproportionate impact on minority candidates.

It is unnecessary for the Court to scrutinize the advisability of incorporating assessments of qualities such as integrity and conscientiousness or the relative merits of the Chicago process. It is sufficient to acknowledge that the existence of such alternative measures and methods belies, as Plaintiffs suggest, Defendants' position that they had no choice but to go forward with the 2002 promotion process despite its adverse impact because no alternative methods with less adverse impact were available.

Defendant argues that Plaintiffs have failed to meet their burden because none of the alternatives now suggested were proposed at the time the 2002 process was implemented. This argument misconstrues the appropriate standard. Plaintiffs must prove that there was "another *available* method of evaluation which was equally valid and less discriminatory." *Bryant v. City of Chicago*, 200 F.3d 1092, 1094 (7th Cir. 2000) (emphasis added). Plaintiffs are not required to have proposed the alternative. The requirement is only that the alternative was available. The Court reads "availability" in this context to mean that Defendant either knew or should have known that such an alternative existed. Plaintiffs have amply demonstrated that Defendant knew of all three alternatives they have set forth.

(R. 388, Bench Trial Op. at 25–26.)

Notably, the court relies on the relative success of the 1996 test, without (1) requiring evidence that the 2002 process would benefit from incorporating the 1996 test's simulation, or (2) addressing the City's interest in test-security, in light of the 1996 simulation's documented cheating. Also, the district court expressly declines to consider the merits of the integrity/conscientiousness and Chicago-plan alternatives, resting its conclusion solely on the City's denial of alternatives.

D. The City's Challenge to the Court's Analysis

The City challenges the district court's judgment, asserting both legal error and factual deficiencies with plaintiffs' step-three showing. Though plaintiffs characterize the City's argument as an attack on the district court's factual findings, invoking the deference of clear-error review, the district court's analysis contains legal errors subject to our de novo review. *Beaven*, 622 F.3d at 547.

First, the district court readily admits crediting the Chicago-plan and integrity/conscientiousness alternatives without considering their relative merit; this approach conflicts with Title VII's requirement that plaintiffs prove the availability of equally valid, less discriminatory measures. *See* 42 U.S.C. §§ 2000e(m), 2000e-2(k)(1)(A)(i)–(ii); 29 C.F.R. § 1607.3(B); *Allen*, 351 F.3d at 316–17; *Shollenbarger*, 297 F. App'x at 487.

Second, the district court accords “considerable significance” to the results of the 1996 simulation with no discussion of the City's test-security concerns. Courts recognize employers' legitimate interest in preserving the integrity of their employment processes. *E.g.*, *Hearn v. City of Jackson*, 340 F. Supp. 2d 728, 742 (S.D. Miss. 2003) (overruling disparate-impact plaintiffs' proposal requiring all applicants to complete a lengthy, interview-based selection procedure, noting the city's legitimate interests in resource preservation, avoiding the appearance of selection bias, and preventing later applicants from obtaining the questions in advance), *aff'd*, 110 F. App'x 424 (5th Cir. 2004) (per curiam).

Here, the City presented undisputed evidence that leaked information and candidate coaching compromised both the 1996 simulation and its 2000-process replacement, a video-based test of law enforcement techniques. (R. 648-6, Trial Tr. (Jones) at 863–65 (discussing the “coaching” problems experienced with the 1996 simulation); R. 648-16, Trial Tr. (Claxton) at 2003 (explaining that City employees were excluded from the creation of the 2002 process, because “city employees are accused of funneling questions and/or answers to participants in a prior process”).) Though candidate coaching did not affect the outcome of the 1996 simulation—evaluators helped poor-performing candidates who would not qualify for promotion—it exposed a security flaw, and the 1996 process's designer testified that the simulator “was [the] weakest link” of the process, noting that “it contributed to most of the race differences” arising from the 1996 process's testing methodologies. (R. 648-7, Trial Tr. (Jones) at 921–22.) The parties certainly knew of these security problems during the development of the 2002 process, as evidenced by Judge McCalla's statements at the parties' June 27, 2001 status conference. (*See, e.g.*, R. 656-17, 6/27/01 Hr'g Tr. at 42 (“[T]he issues that arose in the previous test, we don't want to run the chance of affecting the outcome of the test by giving out unnecessary information”))

Third, the district court's analysis elides the City's concern regarding the impracticability of the 1996 simulation, which required numerous actors to portray the two-hour law enforcement scenarios and took nearly three months to evaluate more than 400 applicants. (*See* R. 648-6, Trial Tr. (Jones) at 863-66.) As the City's expert explained, the protracted nature of simulation testing and the number of moving parts reinforced the City's concerns about testing security. (*Id.*; *see also* R. 648-11, Trial Tr. (Jeanneret) at 1461 (citing "all of the issues that had been raised about the [City's testing] and the confidentiality and . . . prior knowledge of the test and . . . the integrity of the process" as reasons he declined to use the 1996 process).) The court should have accounted for the City's legitimate interests in test security and practicability in assessing plaintiffs' proffered alternatives. *See Watson*, 487 U.S. at 998 (plurality) ("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."); *see also Allen*, 351 F.3d at 314-15 (considering proposal's effect on the city-employer's financial interests); *Clady v. Cnty. of Los Angeles*, 770 F.2d 1421, 1432 (9th Cir. 1985) ("Financial concerns are legitimate needs of the employer."); *Chrisner v. Complete Auto Transit, Inc.*, 645 F.2d 1251, 1263 (6th Cir. 1981) ("Of course, the marginal cost of another hiring policy and its implications for public safety are factors which should not be omitted from consideration.").

Finally, the Seventh Circuit's decision in *Allen* persuades us that the district court erred by relying solely on the past success of the 1996 process in determining that the 2002 process should have incorporated a live simulation. *Allen* similarly involved police officers' challenge to a city's promotion process. The officers proposed eliminating the written job-skills test from the process, so as to give full weight to merit-review boards. *See Allen*, 351 F.3d at 316-17. Noting the absence of "evidence that merit selection is inherently less likely to cause a disparate impact" than the other testing procedures, the court rejected this proposal and affirmed the grant of summary judgment to the city, explaining that "[t]he non-discriminatory history of past merit selection in the [Chicago Police Department] is not sufficient evidence to withstand the City's motion for summary judgment." *Id.* at 317.

In sum, these legal errors improperly shifted plaintiffs' evidentiary burden to the City, undermining the district court's judgment. At a minimum, we must vacate the district court's Title VII judgment. The City asks us to go further, though, and find plaintiffs' step-three showing insufficient as a matter of law. We thus must decide whether plaintiffs' evidence presents a triable issue as to the availability of equally valid, less discriminatory testing alternatives. It does not.

E. Plaintiffs' Insufficient Step-Three Showing

As noted above, the plaintiffs' appellate briefing defends the validity and racial impact of only the 1996 simulation. The plaintiffs first point to the 1996 process's validation report and the City's Answer, which concedes that the 1996 process resulted in no adverse impact. The plaintiffs next highlight their expert's testimony regarding the difference between high-fidelity simulations and the 2002 process's low-fidelity video test. Third, the plaintiffs claim that statistical evidence shows that the 1996 simulation had higher content validity and lower disparate-impact scores than the 2002 process's tests. Finally, the plaintiffs stress the simplicity and affordability of the 1996 process compared to the 2002 process. The scant evidence supporting these claims dooms plaintiffs' reliance on the 1996 simulation as satisfying its step-three burden.

Beginning with the results of the 1996 process as a whole, that evidence does not persuade inasmuch as plaintiffs do not seek to substitute the entire 1996 process for the 2002 process.

As for the expert testimony, plaintiffs' expert, Dr. Richard DeShon, asserted that high-fidelity exercises have greater validity than video-based tests, explaining that law enforcement simulations, like pilot simulators, require the candidate to perform the necessary tasks under realistic conditions. (*See* R. 648-4, Trial Tr. (DeShon) at 533; R. 648-15, Trial Tr. (DeShon) at 1848.⁴) But plaintiffs' briefing offers no data showing that simulations provide equally valid and

⁴We note that Dr. DeShon's initial report in May 2004—more than two years after the administration of the 2002 process—advocated for *both* “role plays and video assessments” as less discriminatory testing methods than written tests. (R. 656-4, DeShon Rpt. at 14.) After Dr. Jeanneret's responsive report alerted him to the 2002 process's inclusion of a video exam (R. 656-5, Jeanneret Resp. Rpt. at 29), Dr. DeShon issued a supplemental report in February 2005 championing high-fidelity simulations, specifically the one used in the 1996 process (R. 656-6, DeShon Suppl. Rpt. at 23).

less discriminatory evaluations than other forms of practical tests.⁵ Moreover, the virtues cited by Dr. DeShon expose another problem with work simulations: scoring subjectivity.

Subjective testing mechanisms open the door to random results and real and perceived scoring bias. *See, e.g., Allen*, 351 F.3d at 315 (“This court previously has noted the potential objection to subjective components of evaluation in selection procedures.”); *Hearn*, 340 F. Supp. 2d at 742 (rejecting panel-interviews proposal, explaining that they “could have contributed to a feeling among candidates that the process was not fair and unbiased”); *Nash v. Consol. City of Jacksonville*, 895 F. Supp. 1536, 1553 (M.D. Fla. 1995) (rejecting subjective performance evaluations, expressing concern that they “would open the process to favoritism, politics and tokenism”), *aff’d*, 85 F.3d 643 (11th Cir. 1996). Tellingly, plaintiffs’ counsel acknowledged this problem during the formulation of the 2002 process when he objected to *the inclusion* of subjective testing components. (*See* R. 657-1, Feb. 26, 2001 Letter to City’s Expert at 4.) Equally revealing, plaintiffs’ appellate briefing remains silent on the subjectivity problem.

We might overlook this pitfall if plaintiffs proffered evidence detailing how a subjective component could be scored so as to minimize disparate impact. But, as discussed, they provide no explanation for how the City should have meshed the 1996 simulation into the 2002 process, whether as a replacement or supplement for the low-fidelity video test, other testing components, or the entire process. Without that type of evidence, plaintiffs lose their argument that use of a high-fidelity simulation would produce better outcomes, because plaintiffs acknowledge that “[e]very single component of the 2002 testing process resulted in ‘very substantial’ adverse impact.” (Third Br. at 34; *see also* First Br. at 23 (detailing the adverse impact of each testing component).)

The plaintiffs likewise neglect to account for the City’s legitimate interests in test security and efficiency. The 1996 simulation, which individually evaluated more than 400 candidates’ law-enforcement techniques via two-hour role-play scenarios, required numerous actors to produce, lasted three weeks, and took two months to grade. (R. 648-6, Trial

⁵Indeed, plaintiffs’ appellate briefing takes inconsistent positions regarding whether a low-fidelity video exam qualifies as a “practical test,” first arguing that it was the essential practical test for the 2000 process, and then arguing that the 2002 process lacked a practical test despite including a video exam. (*Compare* First Br. at 38–39, and Third Br. at 16, with Third Br. at 33.)

Tr. (Jones) at 863–66.) Then the City discovered instances of candidate coaching, for which the plaintiffs prescribe no remedy, seemingly content with their expert’s unqualified assurance that the 1996 simulation would be “easily replicated” at a lesser cost than the 2002 process. (Third Br. at 35 (comparing the costs of the two processes: \$79,250 for 1996, more than \$400,000 for 2002).) But the costs argument overlooks the cheating problems associated with the 1996 and 2000 testing; the City hired outside consultants to design the 2002 process to insulate the exam from the potential biases of City employees. (See Second Br. at 14–15; R. 648-16, Trial Tr. (Claxton) at 2003.) And plaintiffs point to no evidence showing administration of a reliable simulation exercise to more than 500 candidates at a reasonable cost (time and money) and in a manner that minimizes the likelihood of candidate coaching or information leaking. The City’s expert report advised the parties *in 2001* that simulations pose such problems, but when the City proposed a video test at status conferences before Judge McCalla, the plaintiffs expressed no qualms. (See R. 652-4, Jeanneret Rpt. at 38–39; R. 656-17, Status Conf. Hr’g Tr. at 28–32; R. 60, 7/2/01 Status Conf. Order at 1–2; O.A. at 28:10–29:55, 31:50–32:05.⁶)

At bottom, plaintiffs rest their proposal on the actual results of the 1996 simulation, stressing that it produced less racial disparity than the 2002 process. (See Third Br. at 35 (comparing the 1996 simulation’s race-disparity score, $d=.21$, to that of the 2002 process, $d=.83$.) Yet, as the Seventh Circuit explained in *Allen*—and we agree—past practice alone does not suffice. 351 F.3d at 315–17. The “[p]ast success” of a specific testing process “merely predicts, but does not establish, success” in future applications. *Id.* at 315. This broadest of Title VII remedies—which requires no showing of discriminatory motive, see *Griggs*, 401 U.S. at 431—demands evidence that plaintiffs’ preferred alternative would have improved upon the challenged practice. See *Allen*, 351 F.3d at 315 (“We cannot require the City to [incorporate plaintiffs’ alternative testing proposal based] on mere speculation.”); *Zamlen v. City of Cleveland*, 906 F.2d 209, 220 (6th Cir. 1990) (rejecting test-rescoring proposal, where plaintiffs offered only speculation of a less discriminatory impact). This is especially true here, where plaintiffs propose a cumbersome exercise with a track record of security problems, no objective measures of candidate performance, and no explanation for how it could fit into the 2002 process

⁶Though the City’s consultants may not have examined the exact components of the 1996 process, the report and the parties’ discussions before the district court belie the plaintiffs’ claim that the City failed to investigate the possibility of using simulations.

or why it would produce better outcomes. The one-off results of the 1996 simulation, without more, do not carry plaintiffs' burden.

Though arguably forfeited by plaintiffs' minimalist briefing, the Chicago-plan and integrity/conscientiousness-testing proposals fare no better. Again, plaintiffs offer no justification for their comparative validity or discriminatory effect, as compared to the 2002 process's testing features. We further note that the Chicago plan's use of merit-review boards suffers from the same subjectivity and speculation problems identified by the Seventh Circuit in *Allen*. See 351 F.3d at 315–17. As for integrity/conscientiousness testing, EEOC guidelines generally disfavor tests that measure abstract character traits by making inferences about candidates' mental processes. See 29 C.F.R. § 1607.14(C)(1) (“A selection procedure based upon inferences about mental processes cannot be supported solely or primarily on the basis of content validity. Thus, a content strategy is not appropriate for demonstrating the validity of selection procedures which purport to measure traits or constructs, such as intelligence, aptitude, personality, commonsense, judgment, leadership, and spatial ability.”). Plaintiffs acknowledge as much. (Third Br. at 9.) With this in mind, the plaintiffs' expert's vague support for some sort of integrity/conscientiousness testing cannot demonstrate an equally valid, less discriminatory alternative. (See Third Br. at 29; R. 684-13, Trial Tr. (DeShon) at 1681; R. 648-4, Trial Tr. (DeShon) at 670.)

Ultimately, the district court aptly described plaintiffs' proposed alternatives as “broad suggestions.” No doubt, the 2002 process resulted in a substantially higher percentage of unsuccessful African-American applicants. But plaintiffs must offer more to establish a Title VII disparate-impact violation. Because plaintiffs failed to present evidence establishing a genuine issue of fact regarding the availability of equally valid, less discriminatory alternative testing methods, their step-three showing fails as a matter of law.

Perhaps anticipating this outcome, plaintiffs offer an alternative defense of the district court's Title VII judgment that assails the City's step-two showing (credited by the district court) that the 2002 process was job-related and consistent with business necessity. See *Ricci*, 557 U.S. at 578. Accordingly, we backtrack to the step-two standard.

IV. PLAINTIFF'S ALTERNATIVE DEFENSE OF TITLE VII JUDGMENT: THE CITY'S STEP-TWO SHOWING

“Once the plaintiff succeeds in making a *prima facie* disparate-impact case, the defendant may avoid liability by showing that the protocol in question has a manifest relationship to the employment.” *Davis*, 717 F.3d at 494 (citation and internal quotation marks omitted). The City may meet its step-two burden by showing through “professionally acceptable methods, [that its testing methodology is] 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.” *City of Akron*, 824 F.2d at 480 (citation and internal quotation marks omitted). Courts often refer to a test’s job-relatedness and business necessity in terms of its “validity”—denoting the test’s relationship to relevant job content—and “reliability”—referring to its ability to produce consistent results. *See, e.g., Guardians Ass’n of N.Y. City Police Dep’t, Inc. v. Civil Serv. Comm’n*, 630 F.2d 79, 101 (2d Cir. 1980). When the employment position involves public safety, we accord greater latitude to the employer’s showing of job-relatedness and business necessity. *Chrisner*, 645 F.2d at 1262–63 (finding sufficient support for an employer’s truck-driving experience requirements, noting that “[a]n industry with the primary function of managing the safety of large numbers of passengers must be allowed more latitude in structuring the requirements which could [a]ffect the performance of a primary business objective”); *see also Spurlock v. United Airlines, Inc.*, 475 F.2d 216, 219 (10th Cir. 1972) (“[W]hen the job clearly requires a high degree of skill and the economic and human risks involved in hiring an unqualified applicant are great, the employer bears a correspondingly lighter burden to show that his employment criteria are job-related.”).

The City used a “content validity” model for the 2002 process that tests a “representative sample of the content of the job.” 29 C.F.R. § 1607.14(C); *accord Gonzales v. Galvin*, 151 F.3d 526, 529 n.4 (6th Cir. 1998) (citing, as an example of a content exam, a secretary’s typing test). We recognize that a police department’s selection of testing criteria “is largely a matter within the professional judgment of the test writer based upon the particular attributes of the job in question.” *Police Officers for Equal Rights v. City of Columbus*, 916 F.2d 1092, 1099–1100 (6th Cir. 1990) (affirming the district court’s conclusion that job-relatedness “does not require precise proportionality” between the exam content and the relative importance of job tasks).

A. District Court's Validity Findings

Here, in deeming the 2002 process's testing methods valid, the district court detailed Dr. Jeanneret's "comprehensive job analysis," on behalf of the City, to identify the most important knowledge, skills, abilities, and personal characteristics (KSAPs) for the sergeant position.

Jeanneret & Associates sought to assess all 44 of the important KSAPs identified in the job analysis and designed the test questions to meet the content validity requirements for the assessment. The investigative forms and other materials used in the investigative logic test and oral component were very similar to the actual materials used on the job and clearly simulated critical job duties. Additionally, all of the items on the job knowledge test were developed using the same reference materials used by MPD sergeants on the job. The investigative logic test involved realistic scenarios that were designed to simulate situations encountered and investigative activities performed by sergeants on the job. Likewise, the application of knowledge test was designed to evaluate how a candidate would respond to common situations encountered on the job. The [video-based] oral component also involved realistic scenarios designed to simulate situations in which a sergeant would be expected to use oral communication skills in responding to a superior officer, responding to the mother of a victim, and responding to a new partner.

(R. 388, Bench Trial Op. at 17, 19–20.) Other than baldly saying that the tests did not measure traits relevant to the sergeant position (*see* Third Br. at 9)—arguments that appear to circle back to the claim that the 2002 process needed a work simulation instead of the video test—plaintiffs cite no evidence that contests the job-relatedness or representativeness of the KSAPs measured in each test component. We discern no clear error with these validity findings.

B. District Court's Findings Regarding Reliability & Rank Ordering

Plaintiffs devote most of their alternative argument to the district court's findings regarding reliability and rank ordering. On reliability, the court found:

[The City's expert and the designer of the 2002 process] Dr. Jeanneret testified that he did not include a reliability estimate in the validation report because the 2002 process was heterogeneous, i.e., it measured numerous broad KSAP dimensions that were correlated with one another, and he felt that there was no appropriate estimate of reliability. According to Dr. Jeanneret, the most appropriate approach to reliability for such a heterogeneous test was test-retest reliability, which was not feasible under the circumstances. A reasonable alternative, Dr. Jeanneret asserted, would have been to develop an alternate form, requiring two identical tests which, he believed, was not possible in light of the

particular testing environment. Since neither multiple administrations of the test nor parallel administration of identical tests were practicable, Dr. Jeanneret believed the only potentially applicable method of assessing reliability was to measure internal consistency using “coefficient alpha.” Dr. Jeanneret did not initially compute coefficient alpha because he intentionally designed a very heterogeneous test and making coefficient alpha, in his opinion, an inappropriate index of reliability.

Both Dr. Jeanneret and [plaintiffs’ expert] Dr. DeShon subsequently measured coefficient alpha, using somewhat different methodologies. Dr. DeShon reported an overall reliability coefficient of .76 using a method known as stratified alpha. Dr. DeShon included seniority in his analysis, which Dr. Jeanneret testified was inappropriate because seniority was not part of the measurement process. (Jeanneret, Tr. Vol. 11, 1287–88; DeShon, Tr. Vol. 5, 575; Tr. Vol. 16, 1898, 1912.) The Court agrees that inclusion of seniority was inappropriate in assessing the reliability of the test. Since seniority was an administrative add-on component, there is no reason to expect that there would be a significant correlation or internal consistency between seniority and test items. Dr. Jeanneret eventually performed a reliability analysis using a “linear composite,” which resulted in a coefficient of .82. He also computed reliability using the formula for stratified alpha, which resulted in a coefficient of .83.

The Court finds credible Dr. Jeanneret’s testimony as to the limited applicability of coefficient alpha in measuring reliability of a heterogeneous test which draws material for test items from multiple sources. The Court further finds that Dr. Jeanneret’s computations of stratified alpha without inclusion of seniority scores to be more appropriate than Dr. DeShon’s computation, which included seniority. Finally, the Court finds that Dr. Jeanneret’s conclusion that the 2002 process was sufficiently reliable is consistent with professional standards and is supported by relevant law. *See Hearn v. City of Jackson*, 340 F. Supp. 2d 728, 740–41 (S.D. Miss. 2003) (finding that a reliability coefficient of .79 is a common and acceptable value in the context of a heterogeneous test environment).

(R. 388, Bench Trial Op. at 21–22 (transcript citations omitted).)

On the subject of rank ordering, the court found:

Under both Sixth Circuit precedent and the Guidelines, ranking of candidates is appropriate where it can be shown that a higher score correlates with higher job performance. *See Williams v. Vukovich*, 720 F.2d 909, 924 (6th Cir. 1983); 29 C.F.R. § 1607.14(C)(9) (2006). The requirements for rank ordering can be met through a substantial demonstration of job-relatedness, variance in test scores, and an adequate degree of test reliability. *Guardians Ass’n of New York City Police Dep’t, Inc. v. Civil Serv.*, 630 F.2d 79, 104 (2d Cir. 1980).

As discussed above, the test content of the 2002 process was substantially job-related and there was an acceptable level of test reliability. Many sections of the test consisted of items in which there were several right answers, with differing point values for various elements, and/or opportunities for additional credit, all of which serve to distinguish better performing candidates from lesser performing candidates. (Def's Ex. 22, pp. 43–46.) The written test was closely modeled after the like section in the 2000 process, which Dr. DeShon acknowledged was able to differentiate between those candidates with more job knowledge from those with less knowledge. (DeShon, Tr. Vol. 5, 546–47.) Additionally, the raw scores on the 2002 assessment show a substantial variance, with the highest raw score of 358.750 and the lowest of 174.750, among 517 candidates. (Def's Ex. 17.) *See City of Columbus*, 916 F.2d at 1102–03 (upholding rank ordering where score range was 40 points among 71 candidates).

Based on the foregoing, the Court finds that rank ordering of the results of the 2002 process was proper, given that the test had an acceptable level of test reliability, was substantially job-related, and had substantial variance among the scores.

(*Id.* at 22–23.)

Plaintiffs lodge several objections to the reliability and rank-ordering findings, laced with a variety of counter-evidence in the opening of their response brief. (*See* Third Br. at 3–15, 44–62.) We distill three primary arguments: (1) that the district court incorrectly determined that Dr. DeShon incorporated seniority into his composite reliability score, and thus clearly erred in crediting Dr. Jeanneret's reliability testimony; (2) that the district court applied the wrong legal standard for rank ordering, and the City failed to justify rank ordering by showing that higher test scores resulted in better job performance; and (3) that the district court erred by accepting the City's use of seniority in the 2002 process. None demonstrates a reversible legal error or clearly erroneous factual finding.

1. Dr. DeShon's Non-Use of Seniority & the Court's Credibility Finding

First, plaintiffs deny the district court's factual assertion that Dr. DeShon included seniority in his reliability calculations. The City appears to concede the inconclusive nature of the evidence cited by the district court (*see* Fourth Br. at 27–28), but notes that any error in this regard is harmless because both experts' reliability scores (.76 from DeShon, .82–.83 from Jeanneret) fall within the range of reliability scores accepted by courts. *See, e.g., Hearn*, 340 F. Supp. 2d at 740 (approving of exam with .79 reliability coefficient). Yet any mistake regarding

the constituent parts of Dr. DeShon's composite reliability score (.76) leaves undisturbed the court's remaining credibility determinations pertaining to *Dr. Jeanneret's* reliability methodology and testimony—namely, its approval of (1) “Dr. Jeanneret's testimony as to the limited applicability of coefficient alpha in measuring reliability of a heterogeneous test which draws material for test items from multiple sources,” and (2) his “conclusion that the 2002 process was sufficiently reliable.” (R. 388, Bench Trial Op. at 21–22.)

The court's remaining conclusion—choosing Dr. Jeanneret's reliability estimates (.82–.83) over that of Dr. DeShon (.76)—suffers only from the court's mistaken belief that Dr. DeShon's figure included seniority. So far as we can tell, plaintiffs accept the court's related finding that these specific reliability calculations *should not include seniority*. Surprisingly, for all their complaints about Dr. Jeanneret's methods, plaintiffs voice no concern for the higher result he achieved (.82 or .83⁷) using their preferred calculation method, stratified alpha. Arguably, the district court selected Dr. Jeanneret's number because it found his testimony more credible (consistent with its other credibility findings on this issue), not because it believed that Dr. DeShon made a calculation error. And even if the district court chose Dr. DeShon's reliability number (.76), the district court cited authority approving a similar reliability coefficient. *Hearn*, 340 F. Supp. 2d at 740–41 (.79); *cf. Nash*, 895 F. Supp. at 1548 (stating that a reliability coefficient “above 0.70 is considered to be reliable”). Plaintiffs provide no authority compelling the conclusion that either a .76 or .82–.83 reliability score for this type of test fails as a matter of law.⁸

Instead, plaintiffs charge that Dr. Jeanneret conceded the inappropriateness of his own reliability estimate. To the extent plaintiffs suggest that Dr. Jeanneret rejected his own

⁷Plaintiffs suggest in passing that Dr. Jeanneret did not know of “stratified alpha” and did not calculate it. (Third Br. at 52.) But Dr. Jeanneret explained that, though he initially lacked familiarity with the term “stratified alpha,” the “mathematics of the coefficient . . . [are] basically the same” as the “linear composite” figure he calculated. (R. 648-10, Trial Tr. (Jeanneret) at 1285–86.)

We note that the cited evidence appears to invert the coefficient and stratified alpha scores (.83 and .82) noted by the district court and the City's brief, but plaintiffs make no objection on this ground, and we have no reason to believe that the marginal difference between those two scores matters here.

⁸Of course, we do not suggest that a reliability score of .70 suffices for all tests as a matter of law. Reliability determinations depend on the unique circumstances of the testing protocol. We simply acknowledge that this aspect of plaintiffs' reliability argument asks us to determine credibility—something we cannot do. *Harrison v. Monumental Life Ins. Co.*, 333 F.3d 717, 723 (6th Cir. 2003) (“Since we are not free to disregard the district court's credibility assessment, the verdict must stand if [plausible evidence] supports [it.]”).

calculations, they misread his testimony. (*See* R. 648-12, Trial Tr. (Jeanneret) at 1507 (acknowledging that his original report excluded a reliability coefficient, because it would not be an appropriate measure for the test, and stating his belief “that the coefficient alpha or internal consistency index of reliability [would not be] the most appropriate or even really an appropriate index for the reliability of the [2002 process]”).) As the district court noted, Dr. Jeanneret’s testimony explains the difficulty of calculating a reliability coefficient for a heterogenous test—*i.e.*, one consisting of multiple, unrelated components that evaluate multiple tasks and characteristics. (*See* R. 648-10, Trial Tr. (Jeanneret) at 1273–81.) In choosing between the parties’ similar reliability estimates, the district court reasonably credited Dr. Jeanneret’s testimony that the best reliability measures—retesting candidates or administering duplicate tests—were impracticable for a process administered to more than 500 candidates. *See, e.g., Anderson v. City of Bessemer City*, 470 U.S. 564, 573–74 (1985) (“If the district court’s account of the evidence is plausible in light of the record viewed in its entirety, the court of appeals may not reverse it even though convinced that had it been sitting as the trier of fact, it would have weighed the evidence differently.”).

2. Rank Ordering

Next, plaintiffs challenge the district court’s approval of the City’s use of rank ordering to distinguish between the candidates’ scores, arguing that the court misapplied three legal requirements for this scoring method set by this court in *Police Officers for Equal Rights*: (1) sufficient raw score spread (2) composite and component reliability, and (3) reasonable job analysis. Yet, as the City points out, our decision in *Police Officers for Equal Rights* included no such rule; it merely observed that the employer’s expert used those requirements. *See* 916 F.2d at 1102. Our standard states that “[r]anking is a valid, job-related selection technique only where the test scores vary directly with job performance.” *Id.* (quoting *Williams v. Vukovich*, 720 F.2d 909, 924 (6th Cir. 1983)). The EEOC guidelines for content-validity studies support this approach:

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.

29 C.F.R. § 1607.14(C)(9) (emphasis added). The City satisfies this likelihood threshold with “a substantial demonstration of job relatedness and representativeness,” score variance, and an “adequate degree” of test reliability. *See Guardians*, 630 F.2d at 104; *see also Police Officers for Equal Rights*, 916 F.2d at 1100 (explaining that, while a test should “measure important aspects of the job . . . for which appropriate measurement is feasible,” the job-relatedness requirement does not demand that the test “measure all [job] aspects, regardless of significance, in their exact proportions”).

The City’s evidence clears this hurdle.

a. Job-Relatedness

First, the district court found that the City’s consultants conducted a “comprehensive job analysis” to identify the relevant KSAPs for the sergeant position, and that the test components measured relevant job tasks using similar materials to those used on the job and realistic law enforcement scenarios. (R. 388, Bench Trial Op. at 17, 19–20.) As noted above, the plaintiffs present no specific objection to these job-relatedness findings.

b. Score Variance

Second, the district court found “substantial variance” among the promotion scores: of the 517 tested candidates, the 2002 process yielded a raw-score point spread of 184 points between the highest and lowest candidates (358.75–174.75), out of a possible 384.5 points. (*Id.* at 23.) Our review of the exam results reveals no clear error in this finding. (R. 656-23, 2002 Process Exam Results at 1–14.) Nor do we detect clear error in the court’s finding of significant variance. *Cf. Police Officers for Equal Rights*, 916 F.2d at 1102–03 (permitting rank ordering where “[t]here was a spread of more than forty points among 71 test takers,” the highest score was 89.66, and the passing score was 70).

Though plaintiffs stress that only one point separated approximately 30 of the more than 500 candidate scores, that circumstance pales in comparison to the sort of score-bunching found problematic elsewhere. *See Guardians*, 630 F.2d at 103 & nn.19–20 (finding insufficient reliability for rank ordering where nearly 9,000 applicants, or 2/3 of the passing scores, had scores between 94 and 97, out of 110 possible points). Moreover, the focus on promotional

scores here exaggerates the 2002 process's bunching effect, because the same candidates' raw scores ranged between 303 and 341, or 79.0 and 88.7 on a 100-point scale. (*See* R. 656-23, 2002 Process Exam Results at 3–4.) Varying seniority points (1–10) contributed significantly to this purported bunching problem.

c. Reliability

Third, the district court found sufficient test reliability, crediting Dr. Jeanneret's composite reliability scores of .82–.83. Again, we find no clear error with the court's factual findings and no error with its legal conclusion.

Plaintiffs briefly mention that the individual components of the 2002 process received poor reliability scores ranging from .32–.79. Indeed, the relatively low component reliability scores give pause. *See Police Officers for Equal Rights*, 916 F.2d at 1102 (allowing rank ordering where the exam's component tests achieved reliability scores ranging from .85–.97). Though the district court did not make specific findings regarding component reliability scores, plaintiffs point to no authority requiring such findings to sustain a rank-ordering test. *Cf. id.* at 1103 (holding that “the trial court was not clearly erroneous in accepting . . . [expert] testimony . . . on the issue of reliability and rank order scoring” that happened to include a component reliability estimate) (footnote omitted).

“The district judge is entitled in questions of this kind which require expert [statistical] opinion to rely on that opinion.” *Id.* So too here, where the district court relied on Dr. Jeanneret's opinion that the heterogeneous nature of the 2002 process's component tests made reliability coefficients less appropriate measures of reliability than other, impracticable methods, like test/re-test consistency or dual-test administration. (R. 388, Bench Trial Op. at 21–22.) And, as we said, both the plaintiffs' expert and the City's expert attained composite reliability figures greater than .75 regardless of any reliability problems with the component tests.

Still, the plaintiffs argue that the City produced no evidence that the test scores vary with performance so as to justify rank ordering. *See Williams*, 720 F.2d at 924. And, they add, high standard error measurements (SEM +3.64, +10.09 SED) belie the City's claim of reliable test scores, rendering 428 of the 517 candidate scores statistically indistinguishable. Though the

district court's opinion did not specifically address SEM or SED, neither of these claims undermines its finding that the City demonstrated sufficient reliability for rank ordering. With regard to likely test-score/job-performance correlation, Dr. Jeanneret's supplemental report cited published industry principles asserting that "cognitively based selection techniques developed by content-oriented procedures . . . can usually be assumed to have a linear relationship to job behavior." (R. 656-7, Jeanneret Resp. Suppl. Rpt. at 35 (acknowledging that the 2002 process, while not a cognitive-ability test, had cognitive components).) We also note as significant the district court's finding—unchallenged on appeal—that the 2002 process's "written test was closely modeled after the like section in the 2000 process, which Dr. DeShon acknowledged was able to differentiate between those candidates with more job knowledge from those with less knowledge." (R. 388, Bench Trial Op. at 23 (citing R. 648-4, Trial Tr. (DeShon) at 546–47).)

On the topic of SEM, plaintiffs offer no authority explaining why an SEM range of 2.8 (Dr. Jeanneret's corrected estimate calculated during trial) to 3.7, by itself, renders the 2002 process inherently unreliable or trumps other measurements of reliability. They do not show, for instance, the sort of score-bunching and passage-rates deemed problematic by the Second Circuit in *Guardians*. See 630 F.2d at 103 & n.19 (finding unreliable a rank-ordered promotional test with an SEM of 2.4, explaining that the test "was too easy" and resulted in "8,928 applicants, two-thirds of all who passed, [with] bunched [scores] between 94 and 97" out of a possible 110 points).

As for SED, Dr. Jeanneret's supplemental report provides detailed reasons, supported by industry publications, for not relying on this measurement. (See R. 656-7, Jeanneret Resp. Suppl. Rpt. at 34–35.) Specifically, he opposes using large SED bands to equate broad ranges of test scores, explaining that SED bands "are calculated based on the normal probability distribution," meaning that "the further apart two scores are, the more likely those scores are to be truly different." (*Id.* at 34.) He elaborates, citing an industry publication finding that "even when a test is quite reliable, a typical SED band covers so large a part of the test score range that the preferred interpretation of banding advocates . . . is false." Dr. Jeanneret goes on to note that "test score bands . . . try[ing] to account for measurement error . . . [are] not required, or even

endorsed by the professional standards in the field of industrial and organizational psychology (i.e., *Principles*, 2003; *Standards*, 1999).” (*Id.*)

Ultimately, the district court heard the parties’ competing evidence regarding reliability, SEM, and SED, and the court found that the City justified the use of rank ordering with a substantial demonstration of job-relatedness, score variance, and an adequate degree of reliability supporting the likelihood that test scores would correlate to job performance. We find no clear error with the court’s findings of fact in this regard and no error with its ultimate legal conclusion regarding rank ordering.

3. *Seniority Scoring*

Last, plaintiffs denounce the City’s use and weighting of candidates’ seniority—an item included in their Memorandum of Understanding (MOU) with the officers’ union—as a promotional factor. The Supreme Court has held that a “bona fide seniority system [is not] unlawful under Title VII,” even though “a seniority system inevitably tends to perpetuate the effects of pre-Act discrimination.” *Int’l Bhd. of Teamsters v. United States*, 431 U.S. 324, 352–53 (1977) (construing 42 U.S.C. § 2000e-2(h)). Thus, this court will sustain the seniority component of a promotional procedure “so long as an intent to discriminate did not enter into its adoption and it has been maintained free from any illegal purpose.” *City of Akron*, 824 F.2d at 481.

Though not quarreling with this standard, plaintiffs challenge the binding effect of the MOU on the City. But, contractual enforceability aside, without showing discriminatory intent or illegal purpose, plaintiffs have no grounds to impugn the City’s use of seniority. As for weighting, the plaintiffs suggest that the City’s scoring errors inflated seniority’s impact from an intended 10% to 25%. The cited testimony, however, appears to refer to something other than a tabulation error; Dr. DeShon differentiates between a “nominal weight” of 10% and an “effective” or “actual weight” of 25%, referring to the degree to which seniority affected promotion score variance. (R. 648-14, Trial Tr. (DeShon) at 1753–55.) Review of the test results (raw scores, scaled scores, and promotion scores) confirms this, revealing that seniority accounted for up to 10 points of the promotion score, out of a possible 110 points. (*See generally* R. 656-23.) Regardless of the nature of the alleged scoring error, in the absence of

evidence that the City's weighting of seniority reflects a discriminatory intent or other illegal purpose, plaintiffs gain no ground. *See City of Akron*, 824 F.2d at 481. Because the seniority component required no additional validation, the district court properly rejected this aspect of the plaintiffs' challenge.

V. CONCLUSION

For these reasons, we affirm in part and reverse in part the district court's judgment. We AFFIRM the district court's immunity-based dismissal of plaintiffs' negligence claim related to the 2000 process, but we REVERSE the district court's Title VII judgment invalidating the 2002 process, thereby MOOTING plaintiffs' challenge to the district court's choice of remedies for the 2002 process. We VACATE the district court's fees award and REMAND for further consideration in light of these developments.