In the

United States Court of Appeals

For the Seventh Circuit

Nos. 14-3783 & 15-2030

STACY ERNST, et al.,

Plaintiffs-Appellants,

v.

CITY OF CHICAGO,

Defendant-Appellee.

Appeals from the United States District Court for the
Northern District of Illinois, Eastern Division.
No. 1:08-cv-04370 — **Charles R. Norgle**, *Judge*.

ARGUED FEBRUARY 25, 2016 — DECIDED SEPTEMBER 19, 2016

Before BAUER, MANION and KANNE , *Circuit Judges*.

MANION, *Circuit Judge*. After Stacy Ernst and four other women applied unsuccessfully to work as Chicago paramedics, they brought this Title VII gender-discrimination lawsuit against the City of Chicago. These women were experienced paramedics from public and private providers of emergency medical services; they sought employment as paramedics with the Chicago Fire Department, but they did not apply to

firefighting positions. All five women were denied jobs because they failed Chicago's physical-skills entrance exam.

In district court, this Title VII case was split into two parts. The plaintiffs' disparate-treatment claims went to a jury trial, in which the district court provided an erroneous jury instruction. Their disparate-impact claims were tried in a separate bench trial. This second group of claims turned largely on whether Chicago's test was based on a statistically validated study of job-related skills. We remand for a new jury trial on the disparate-treatment claims, reverse the bench trial's verdict on disparate impact because the physical-skills study was neither reliable nor validated under federal law, and affirm the evidentiary rulings below.

Background

The Chicago Fire Department employs several hundred paramedics.[1] When hiring new paramedics, Chicago has not always tested its applicants' physical skills. From the 1970s through the year 2000, paramedics were hired without any physical test. The hiring process changed in 2000, however, when Chicago implemented a physical-skills test created for it by Human Performance Systems, Inc. Deborah Gebhardt, the president of HPS, led this test-creation process.

Gebhardt had previously created a physical test for the Chicago Fire Department's entry-level firefighters. That test had a disparate impact on women. The plaintiffs argue that Chicago's decision to rehire Gebhardt for the paramedic test,

[1] The factual statements in this opinion are drawn from the record presented on appeal; they should not be read as binding factual findings when the plaintiffs' disparate-treatment claim is retried before a jury.

without taking bids from anyone else, reflects Chicago's desire to reduce the number of women it hired as paramedics.

In this case, Gebhardt tested volunteer Chicago paramedics. These were incumbent paramedics working for the Chicago Fire Department. Gebhardt tested these study volunteers on physical skills designed to reflect job-related skills. She tested the paramedics on three "work samples" also designed to reflect job-related skills. Then she compared the results from the skills testing with the results from the work-sample testing. Through this process, Gebhardt selected physical skills that, together, formed Chicago's physical-skills entry exam for paramedic applicants. This was a concurrent validation study, as this opinion will later explain.

Between 2000 and 2009, nearly 1,100 applicants took Gebhardt's entrance examination. Among these, 800 were men, and 98% of the male applicants passed. Another 300 were women; 60% of female applicants passed. Stacy Ernst, Dawn Hoard, Katherine Kean, Michelle Lahalih, and Irene Res-Pullano took the test in 2004, as licensed paramedics with experience working in other public fire departments or for private ambulance services. In their daily work, they moved patients and did so safely. When they took the Chicago physical-skills examination, however, they all failed.

After they were denied employment based on their exam results, Ernst and her fellow plaintiffs filed suit. They challenged the skills test as discriminatory; they urged that there was no evidence of Chicago paramedics ever lacking the physical ability to properly care for their patients. Instead, they argued, the test was implanted to keep women out. Ultimately, their suit had two parts. On their disparate-treatment

claims, they asked a jury to find that Chicago had a discriminatory motive against women when Chicago implemented its skills test. On their disparate-impact claims, the plaintiffs argued in a bench trial that improper statistical methods were used to establish the skills test.

The jury instruction on disparate treatment was vigorously debated before both the magistrate judge and the district judge. The plaintiffs urged that their burden on this disparate-treatment claim was to prove illegal purpose: that Chicago had a discriminatory intent or motive for implementing the skills test. When arguing before the magistrate judge, Chicago claimed that the plaintiffs had to satisfy a but-for test: that Chicago would have hired the plaintiffs if, all other factors being equal, they were male. In responding to this effort, the magistrate judge said, "That is absolutely not what this case is about at all. At all. And you know it."

The disparate-treatment jury instruction, labeled Jury Instruction 24 at all times in this case, included this language when the magistrate judge settled the instructions:

> Plaintiffs contend the City discriminated against them on the basis of sex in violation of Title VII of the Civil Rights Act of 1964, as amended. In order to succeed on this claim, Plaintiffs must prove by a preponderance of the evidence that the City intentionally created or used the physical abilities test for the purpose of excluding females or reducing the number of females who would be hired as paramedics by the Chicago Fire Department. The City denies that it intentionally created or used the physical abilities test to discriminate against female applicants.

> It is not enough for Plaintiffs to prove merely that the City knew the physical abilities test would have an adverse impact on female applicants. An adverse impact exists where the rate at which female applicants pass the test is substantially less than the rate at which male applicants pass. The parties do not dispute that the test had an adverse impact.

As approved by the magistrate judge, Jury Instruction 24 went on to explain that the plaintiffs should prevail if they prove by a preponderance of the evidence that Chicago "intentionally created or used" the skills test to "exclude or reduce" the women hired as paramedics. If the plaintiffs did not prove this, however, Chicago must prevail. There was no problem with the jury instruction on disparate treatment as established at this point in the litigation.

Chicago was not done urging the but-for test, however, and the City successfully resurrected its argument before the district judge. After a hearing on the matter, the district judge issued a written order that ruled for the defense. He stated that "[b]ecause this is 'an individual action, rather than a class action, evidence of a pattern of practice can only be collateral to evidence of specific discrimination against the plaintiff[s].'" App. 13 (citing *Matthews v. Waukesha Cty.*, 759 F.3d 821, 829 (7th Cir. 2014)) (quotation marks in original). Given this reliance on the individual-action analysis, the district judge struck the original contents of Jury Instruction 24. He inserted the pattern instruction on General Employment Discrimination, so that Instruction 24 now read:

> Each Plaintiff claims that she was not hired as a Chicago Fire Department Paramedic because of her

> gender. To succeed on this claim, each Plaintiff must prove by a preponderance of the evidence that she was not hired by the City of Chicago because of her gender. To determine that a Plaintiff was not hired because of her gender, you must decide that the City would have hired the Plaintiff had she been male but everything else had been the same.

When the case went to the jury, the jurors expressed confusion over this instruction. After deliberating for 90 minutes, they sent a note to the district court: "Question to the Judge regarding instruction 24. Please provide clarification of the sentence, quote: To determine that a plaintiff was not hired because of her gender, you must decide that the City would have hired the plaintiff had she been male but everything else had been the same." While the district court and parties were discussing this, a second note came out: "The jury cannot deliberate further without a response to our question. May we know what the time [is] for a response?" The district court provided this written response, to which the plaintiffs objected: "Reread all instructions. The sentence you are asking to clarify speaks for itself." Four minutes later, the jury returned a verdict for the defense.

During the bench trial on disparate impact, the district court found it clear that the plaintiffs had established a disparate impact on women. The burden therefore shifted to Chicago, which had to prove that its physical-skills test was job-related and consistent with business necessity. In adopting Chicago's proposed conclusions of law, the district court concluded that Gebhardt's validation study satisfied Chicago's burden. The issues at this point in this trial turned on whether

Gebhardt's study satisfied the law's technical standards for validity studies, which appear at 29 C.F.R. § 1607.14(B)(4). The district court wrote that the "[p]laintiffs' arguments attacking Dr. Gebhardt's job analysis, validation study under the criterion method, and the process of determining the 935 passing score are unavailing and are rejected." Dist. Ct. Docket 604 at 2. Accordingly, the burden shifted back to the plaintiffs, who had to show that Chicago had rejected a substantially equally valid, but less discriminatory, alternative to the skills test. On this, the district court concluded that the plaintiffs offered assertions without evidence. The district court thus entered judgment for Chicago.

The plaintiffs lost both trials. They now bring this appeal, which centers on three issues. First, they challenge the disparate-treatment jury instruction that the district judge gave in the jury trial. Second, the bench trial on disparate impact yielded a defense verdict on the statistical methods underlying the skills test, and the plaintiffs challenge those methods. Third, the plaintiffs argue cumulative error from a series of evidentiary rulings. We address these issues in turn.

Discussion

The Civil Rights Acts of 1964 and 1991, known collectively for our purposes as Title VII, prohibit two types of discrimination. *Ricci v. DeStefano*, 557 U.S. 557, 577 (2009). First, Title VII prohibits job-related actions that are motivated by intentional discrimination against employees, based on protected employee statuses such as race or sex. *See id.* (quoting 42 U.S.C. § 2000e-2(a)(1)). This is known as disparate treatment. Plaintiffs must prove that an employer had a discriminatory motive for taking a job-related action. *Id.* During the jury trial in this case, the plaintiffs argued that Chicago adopted its

physical-skills entrance exam in an effort to reduce or eliminate the number of women it hired as paramedics.

Second, Title VII prohibits employment practices that have a disproportionately adverse impact on employees with protected characteristics, even if the impact is unintended. *See id.* This is disparate impact. Employers can defend against a disparate-impact claim by demonstrating that the challenged practice is job-related for the employee's position and consistent with business necessity. 42 U.S.C. § 2000e-2(k)(1)(A)(i). Even if the employer establishes this, however, an employee can still prevail by proving that the employer has rejected an available alternative job practice that (1) results in a less disparate impact, and (2) serves the employer's legitimate needs. *Ricci*, 557 U.S. at 578 (citing 42 U.S.C. §§ 2000e-2(k)(1)(A)(ii), (C)). Chicago does not dispute that its skills test has an adverse impact on women. As Chicago admits, the passing rate for women is about 60% of the passing rate for men. The parties dispute, however, whether the test is job-related and consistent with business necessity.

When reviewing claims that a business practice is insufficiently related to a business necessity, we bear in mind that "'[c]ourts are generally less competent than employers to restructure business practices, and unless mandated to do so by Congress they should not attempt it.'" *Watson v. Fort Worth Bank & Trust*, 487 U.S. 977, 999 (1988) (quoting *Furnco Constr. Corp. v. Waters*, 438 U.S. 567, 578 (1978)).

A. The Disparate-Treatment Claims: Jury Instruction 24

The plaintiffs begin by appealing the jury instruction on their disparate-treatment claims. We review jury instruction challenges de novo. *Lewis v. City of Chi. Police Dep't*, 590 F.3d

427, 433 (7th Cir. 2009). District courts have substantial discretion in how to precisely word jury instructions, provided that the final result, read as a whole, is a complete and correct statement of the law. *Id.* We only reverse when jury instructions are so misleading or confusing that they prejudice a party. *Id.* Though we strive for common-sense readings of jury instructions, and we avoid nitpicking, we also recognize the importance of getting jury instructions right. *See id.* Even when a party fails to object to a jury instruction, a mistaken instruction is preserved for plain-error review if it impacts "substantial rights." Fed. R. Civ. P. 51(d)(2).

In giving Jury Instruction 24, the district judge relied on our ruling in *Matthews v. Wakesha County*, 759 F.3d 821 (7th Cir. 2014). That case is distinguishable. There, Bernadine Matthews applied for employment as an Economic Support Specialist or Supervisor in Waukesha County, Wisconsin. Because she did not satisfy the minimum requirements for either job, she never received a job offer. Matthews eventually sued Waukesha County for racial discrimination. She argued that she should be allowed to bring a Title VII disparate-treatment claim based, not on evidence of racial discrimination against her in particular, but on statistics indicating that racial discrimination exists against blacks in general. This is the argument we rejected in *Matthews*, where we explained that she "would need to present evidence indicating that racial discrimination was the employer's standard operating procedure—the regular rather than unusual practice." *Id.* at 829. This is why we stated that "'evidence of a pattern or practice can only be collateral to evidence of specific discrimination against the plaintiff.'" *Id.* In *Matthews*, the plaintiff failed to present any disparate-treatment claim at all.

In contrast, the plaintiffs in this case argue that Chicago created a new standard operating procedure, with the specific intention of reducing or removing women from among its new paramedic hires. They do not rely on generalized claims of statistical bias against women; instead, they argue that there was no legitimate professional or safety need for Chicago to implement this particular skills test. These arguments place Ernst and her fellow plaintiffs in a different category than Matthews. Whether or not they should win, they at least presented a proper disparate-treatment claim, as indicated by the district court's denial of summary judgment. At trial, the plaintiffs also presented enough evidence to at least support a correct instruction on disparate treatment.

Here, the jury should have been instructed on the plaintiffs' burden of proving that Chicago was motivated by anti-female bias, when Chicago created the entrance exam that caused these plaintiffs not to be hired. *See Ricci*, 557 U.S. at 577. Instead, jurors were instructed on a different burden, which failed to address Chicago's motive for creating the skills test: "To determine that a Plaintiff was not hired because of her gender, you must decide that the City would have hired the Plaintiff had she been male but everything else had been the same." This instruction focused on gender as a factor in the specific decisions not to hire these five plaintiffs, without expressly stating the mandatory question: whether Chicago had an anti-female motivation for creating its skills test. The magistrate judge's version of Instruction 24 more accurately reflected Title VII's focus on whether there was a discriminatory motive behind Chicago's conduct. *See id*.

This legal error would be enough to establish prejudice, but the record goes a step further. It shows that the jurors saw

this instruction as the pivotal issue before them, particularly when they sent a note stating that "[t]he jury cannot deliberate further without a response to our question." Only four minutes after the district judge instructed them to take Instruction 24 at face value, they returned a defense verdict. Under these circumstances, we must remand the disparate-treatment claims for a new trial with proper instruction, namely, the magistrate judge's version of Jury Instruction 24.

B. The Disparate-Impact Claims: Validating the Skills Test

Having addressed the jury instruction on disparate treatment, we turn to disparate impact. The district court found no problem with Gebhardt's job analysis and validity study. It thus entered a verdict for the defense. Because the disparate-impact rulings were made in a bench trial, we review the district court's legal conclusions de novo and review factual findings for clear error. *Bridgeview Health Care Ctr., Ltd. v. Clark*, 816 F.3d 935, 937–38 (7th Cir. 2016).

To prove a disparate-impact case, a plaintiff must show an adverse impact on employees with a protected characteristic like gender. Chicago concedes that its physical-skills entrance test has an adverse impact on women. The burden thus shifts to Chicago, which must show that its physical-skills testing is job-related for the employee's position and consistent with business necessity. 42 U.S.C. § 2000e-2(k)(1)(A)(i). In this case, Chicago relies on Gebhardt's validity study to establish that its physical-skills test is job-related. Employers are not required to support their physical-skills tests with formal validation studies, which "show[] that particular criteria predict actual on-the-job performance." *Watson*, 487 U.S. at 998. When

an employer relies on a validity study, however, federal regulations establish technical standards for these studies. *See* 29 C.F.R. § 1607.14(B)(4).

1. A technical explanation

Before we analyze the federal regulations on validity studies, some technical explanation may be useful. We begin this section by discussing the terms used in federal regulations and Gebhardt's validity study, in an effort to help readers navigate the complex record we are examining.[2]

Validity is the extent to which a study accurately measures what it sets out to measure. In this case, Gebhardt's study sought to measure the physical skills of incumbent Chicago paramedics. Her study was valid to the extent that it accurately measured their physical skills.

The type of validity study that Gebhardt chose was a criterion-related validity study. Researchers decide what criteria they use. Here, Gebhardt solicited job-performance ratings from the volunteer paramedics' supervisors and peers. This was one set of criteria. In addition, Gebhardt created work

---

[2] We recognize that the court is not a statistical expert, yet we must also acknowledge that the law mandates statistical discussion in this case. We therefore note that the terms described here are drawn from federal regulations, the parties' briefs, trial testimony, Gebhardt's statistical report as it appears in the record, and other studies. Our use of technical terms is also consistent with the use of these terms in previously issued court rulings. *See*, *e.g.*, *Watson*, 487 U.S. at 998 ("formal 'validation studies' show[] that particular criteria predict actual on-the-job performance."); *Guardians*, 633 F.2d at 244 ("Criterion-related validity studies correlate test scores with job performance.").

samples, which were supposed to represent on-the-job skills. The work-sample scores were another set of criteria.

A criterion-related validity study measures a study's validity by comparing the assessment-tool results with the criteria. In this case, the assessment tool is a test of physical skills. Gebhardt would correlate the results of the assessment tool (the skills-test scores) with the results of the criteria (either the job-performance ratings or the work-sample scores). If there is a strong correlation, the assessment tool is validated. From a validated study, Gebhardt could conclude that her skills tests accurately assess test-takers for whether they have the physical skills that paramedics learn on the job. In contrast, if there is a weak correlation or no correlation at all, the assessment tool is not validated. From a study that is not validated, Gebhardt would not conclude that her skills testing could assess actual skills learned on the job.

There are two types of criterion-based validity studies: a researcher can conduct a predictive validity study or a concurrent validity study. The predictive approach may be somewhat stronger, as suggested by the fact that federal regulations require courts to conduct additional scrutiny into concurrent validity studies. *See* 29 C.F.R. § 1607.14(B)(4).

In statistical terms, the difference between predictive and concurrent validity is simply a matter of timeline. College entrance exams like the SAT are examples of predictive studies. In this model, the SAT is administered to high school students, the students who took the SAT attend college, and then their college GPAs are examined. If there is a significant correlation between the students' SAT scores and their college GPAs, the SAT test is considered valid. In this sense, the SAT scores are tested against the valid college GPAs.

Here, Gebhardt chose to conduct a concurrent validity study when she tested Chicago's volunteer paramedics and created a physical entrance exam. In a concurrent study, the researcher takes two measures at the same time. The researcher then uses one measure (which is known to be valid) to validate the other measure (which needs to be validated).

First, Gebhardt measured the volunteers' physical skills by having them perform physical skills that she determined were necessary to the paramedic job: a modified stair-climb, leg lifts, arm-strength tests, and other tests. Gebhardt's volunteer paramedics had higher scores than the scores of other paramedics in both public-sector and private-sector jobs. The men in her study could handle an average of 281.9 pounds in leg-lift tests, for example, while the men in a study of several hundred paramedics could handle an average of 245.11 pounds in leg-lift tests. Gebhardt stated that this disparity between volunteers in her study and volunteers in other studies was "especially" true between the tested female paramedics. In an effort to soften the Chicago paramedics' unusually high scores, Gebhardt added scores from another physical test of New York City paramedics. She only used the New York City data, however, when setting a passing score. She did not use it to validate the Chicago study.

Second, Gebhardt also created a rating instrument that she distributed to the volunteers' supervisors and peers. Because these volunteers were incumbent Chicago paramedics, she could obtain on-the-job assessments of these volunteers' abilities from supervisors and peers who also worked in the Chicago Fire Department. On their own, these ratings would be a valid assessment of the volunteers' job skills. Thus, Gebhardt could compare these supervisor and peer ratings

with her skills tests. If the ratings and the test scores yielded comparable assessments of the volunteer paramedics, then Gebhardt could conclude that her skills tests were validated.

The job-performance ratings and the skills-testing scores, however, yielded significantly different assessments of the volunteer paramedics' on-the-job abilities. Based on supervisor and peer ratings, female paramedics' performance was not far from male paramedics' performance: the average female rating was 90% to 93% of the average male rating. If the skills tests were validated by the job-performance ratings, they would have yielded similar results, with no great discrepancy between female and male skills scores. But in Gebhardt's skills test, women performed far less well than men. On the leg-lift test, for example, the average female score was 66.4% of the average male score. This discrepancy would appear to actually invalidate the physical-skills tests.

Rather than setting aside her original skills tests, and creating a new set of tests that might better assess paramedic job skills, Gebhardt provided rationales for setting aside the job-performance ratings. Gebhardt received ratings for 46 out of 52 volunteer participants. In a study-planning letter to the Chicago Fire Department, Gebhardt's company had stated that a minimum of 110 participants would be necessary to validate this study. Though she had been willing to drop from 110 volunteers to 52, she concluded that she could not go from 52 to 46. Further, when Gebhardt set aside the supervisor and peer ratings, she replaced them with work-sample scores, as this opinion will soon explain. But when Gebhardt tested the reliability of her work samples, she had only 7 volunteers in the stretcher lift, 17 volunteers in the stair-chair push, and 18 volunteers in the lift and carry. For work samples, she was

comfortable relying on a far smaller number than 46 volunteers. This calls into question whether going down to 46 volunteers, with the supervisor and peer ratings, was really the problem in the researcher's mind.

In addition, Gebhardt said that she would have had to drop the modified stair-climb component of her skills test if she accepted the supervisor and peer ratings, and women performed better on this stair-climb than on other skills. This created the appearance of favoring women, by preserving an aspect of the test on which they did well. But because the supervisor and peer ratings also did not validate other skills tests, she would also have had to drop skill tests on which women performed worse than men. The leg-lift, for example, was not validated by the job-performance ratings. The average female score on that skills test was 66.4% of the average male score. Yet Gebhardt ultimately included this skill on Chicago's entrance exam. Thus, though Chicago only claims that Gebhardt dropped the job-performance ratings in order to preserve skills testing on which women performed well, the record conflicts with that narrow version of the facts.

Regardless of the reasons for setting aside the job-performance ratings, Gebhardt still needed a concurrent measure to validate her skills tests. Thus, she compared the results of the skills tests with the results of her work-sample tests. She designed three work samples with input from the Chicago Fire Department: a lift and carry, a stair-chair push, and a stretcher lift. These work samples were intended to reflect skills that Chicago paramedics learn on the job.

In the lift and carry, a volunteer lifted a piece of equipment, carried it up a set of stairs, put it down, lifted another piece of equipment, carried that down the stairs, and then put

that down. This required five timed cycles, with faster times resulting in better scores. In the stair-chair push, the volunteer navigated a stair chair over a ramp, with a dummy seated in the stair chair. Again, faster times resulted in higher scores. In the stretcher lift, volunteers lifted a stimulated stretcher to an arm-locked position, held it for 20 seconds, rested for five seconds, and repeated. The stretcher weighed 90 pounds with the first lift, and 10 pounds was added each time, up to a maximum of 220 pounds. This test continued until the volunteer completed 13 cycles or could no longer lift the stretcher. Volunteers did not receive higher scores for performing this more quickly. Instead, scores were based on two measures: cycles completed and weight lifted.

When Gebhardt examined the correlation between the skills tests and the three work samples, she found that three of the skills were validated: the modified stair-climb, arm-endurance test, and leg lift.[3] The correlation between these three skills and the three work samples exceeded the .01 level of statistical significance. This means there was a 99% probability that the results would repeat if they were tested again—a statistically rigorous result that exceeds the .05 level required by law. *See* 29 C.F.R. § 607.14(B)(5).

In addition to testing the validity of her study, Gebhardt tested the reliability of her study. *See* 29 C.F.R. § 1607.14(C)(5). An assessment tool is considered reliable if it produces consistent results over time. Gebhardt chose the test-retest reliability approach, in which a researcher administers a test and then readministers it. If there is a strong correlation between

---

[3] The remaining skills were not validated by the work samples. Gebhardt set those skills aside.

the test and retest, the assessment tool is considered reliable. In technical terms, this process yields what is called a reliability coefficient: a number between 0 and 1. If the reliability coefficient is 1, there is a perfect correlation between the test and the retest. This suggests that the assessment tool is highly reliable because it produces highly consistent results over time. Perfectly repeating results, however, are unusual. If the reliability coefficient is 0, there is no correlation between the test and the retest. This would indicate that the assessment tool is not reliable at all.

In this case, Gebhardt provided test-retest reliability results for the lift and carry (measured by seconds), the stair-chair push (measured by seconds), the stretcher lift (measured by weight), and the stretcher lift again (measured by cycles completed). For the lift and carry, the reliability coefficient was a mere 0.503, indicating about a 50/50 chance that this test is reliable. For the stair-chair push, the reliability coefficient was a moderate 0.743. For the stretcher lift as measured by weight, the reliability coefficient was a robust 0.982. And for the stretcher lift as measured by cycles, the reliability coefficient was 0.978, also indicating strong reliability.

Based on this work by Gebhardt, Chicago implemented a physical entrance exam with three components: the modified stair-climb, arm-endurance test, and leg lift. The passing score was set with this formula, which favored the modified stair-climb on which women did well: (7 · modified stair-climb score) + (2 · arm-endurance score) + (1 · leg-lift score).

2. An analysis of the validity study

With that technical explanation, we turn to our analysis. We consider whether the district court was correct in finding

that Gebhardt's job analysis and physical-skills study satisfied the express federal regulations on validity studies.

In the Title VII context, a validity study examines whether an employer is using an appropriate selection procedure, like Chicago's physical entrance examination, in its hiring process. *See* 29 C.F.R. § 1607.5(B). Federal regulations require that the validity study must establish specific criteria, which empirically demonstrate that the selection procedure predicts or significantly correlates with important job-performance elements.[4] 29 C.F.R. § 1607.5(B). In this case, the specific criteria are the physical skills that Gebhardt tested against work samples, to see whether the skills could be validated as job-related skills. She ultimately found three valid.

The technical standards for validity studies are set forth in 29 C.F.R. § 1607.14(B)(4). This section requires that the volunteers in a study's sample population should, as far as possible, "be representative of the candidates normally available in the relevant labor market for the job." 29 C.F.R. § 1607.14(B)(4). Representativeness is a fact-sensitive inquiry.

Section 1607.14(B)(4) provides two specific guidelines for determining whether a sample population is representative.[5]

---

[4] These criteria are deemed "relevant to the extent that they represent critical or important job duties, work behaviors or work outcomes as developed from the review of job information." 29 C.F.R. § 1607.5(B)(2). The possibility of bias must also be considered when choosing and applying these criteria. 29 C.F.R. § 1607.5(B)(2).

[5] These are not the only legal issues that Section 1607.17(B)(4) requires courts to consider, but we conclude that these are the only issues that apply in the case before us. Gebhardt added scores from a New York City paramedic study to the scores from her Chicago paramedic study. If she used the New York data to help validate the Chicago data, we would be

First, we must consider whether the individuals in the sample population (here, the volunteer incumbent paramedics) are representative of individuals who are normally available in the Chicago paramedic market. As far as possible, the sample population should also include the races, sexes, and ethnic groups normally available in that job market.

Second, in a concurrent validity study like this one, we must examine whether the test focuses on specific skills or knowledge that are the "primary" focus of skills or knowledge that Chicago paramedics learn on the job. 29 C.F.R. § 1607.14(B)(4). We note that Title VII regulations do not provide for employers to implement a general physical *fitness* test. Instead, Title VII regulations look for a physical *skills* test. These skills must specifically relate to skills that Chicago paramedics learn in their jobs. Of course, physical skills may bear a close relationship to physical fitness, but the entry exam may not merely examine generalized strength.

With these federal regulations on validity studies, we first examine whether the volunteer paramedics in Chicago's study are representative of individuals who are normally available in the Chicago job market. *See* 29 C.F.R. § 1607.14(B)(4). The plaintiffs object to the fact that Gebhardt

---

required to conduct an additional inquiry into questions like whether the samples are comparable in terms of actual jobs performed. 29 C.F.R. § 1607.14(B)(4). There is no indication, however, that she used the New York data to validate the Chicago study results. Instead, it appears that she only combined the New York and Chicago data for the purpose of setting a passing score on the final physical-skills test. And we do not reach the question of whether Gebhardt set an appropriate passing score. Thus, we do not conduct additional inquiries into the other issues raised by Section 1607.17(B)(4), such as actual jobs performed in New York.

asked existing Chicago paramedics to volunteer, rather than randomly selecting study participants. This self-selection presents an obvious concern: when an employer asks its employees to volunteer for testing, the strongest employees are most likely to volunteer. Any study results may thus be skewed, instead of representing the general population. Yet as Gebhardt explained, people cannot be forced into studies.[6] There are ways to see that, even with volunteer participants, a study offers legitimate insights into the general population. On its own, the fact that Gebhardt worked with volunteers is not a basis for setting aside her study results.

By Gebhardt's own testimony, however, these volunteer paramedics did not represent the skill-set in the general population of Chicago paramedics. Gebhardt testified that the Chicago volunteers performed better than public-sector and private-sector paramedics normally perform. In her formal re-

---

[6] The researcher's method of selecting volunteers will have significant implications for how representative a sample population is, and the record does not address how Gebhardt's volunteers were chosen, but human participants must be willing participants in any event. This is a principle to which the sciences, social and otherwise, have hewn closely since the Belmont Report was issued in 1979. The Belmont Report was largely a response to reports that people were abused in biomedical experiments during the Second World War. It established three major principles for studying humans: respect for persons (protecting individual autonomy), beneficence (promoting individual wellbeing), and justice (rendering what individuals deserve). Under these guidelines, when a person participating in a study is capable of giving informed consent, the scientist conducting the experiment should not proceed without it. *See* Nat'l Comm'n for the Protection of Human Subjects of Biomedical & Behavioral Research, *The Belmont Report* (Apr. 18, 1979), http://www.hhs. gov/ohrp/regulations-and-policy/belmont-report/index.html#xbasic.

port on this study, Gebhardt forthrightly stated that "the Chicago EMS personnel were found to be above average. This was especially true for the women."

Because she was concerned that the results of her Chicago study might not be representative, she combined her data on 52 Chicago paramedics with a comparable data-set on 87 New York City paramedics. As Gebhardt's study report stated, and Chicago affirms in its brief, she did this "[t]o avoid artificially inflating the passing score." Appellee Br. at 10. In other words, the New York paramedics presumably had lower scores than the Chicago paramedics, which helped draw the average score toward a more normal performance level. For the combined Chicago and New York City scores to result in a truly normal or average score, however, the New York City paramedics' scores would have to be significantly lower than normal. There is no evidence—nor would we expect—that the New York City paramedics perform at a lower skill level than paramedics who are "normally available in the [Chicago] labor market." 29 C.F.R. § 607.14(B)(4).

With 52 Chicago paramedics and 87 New York City paramedics, Gebhardt had a total sample population of 139 paramedics. And with this small sample size, the 52 Chicago paramedics constituted 37% of the total sample population. About four in ten paramedics, in the overall sample population, were still from Chicago. This is still a sufficiently high percentage for Chicago paramedics to pull up the average scores and skew the overall results. Chicago does not explain why it believes that, by adding 87 New York scores, the problem of abnormally high scores was resolved. Perhaps this is a problem that could be resolved by comparing the combined Chicago and New York results with the studies that Gebhardt

considered, when she concluded that the Chicago volunteer paramedics had abnormally high scores, and showing that the combined results are comparable. But the record does not indicate that even the combined sample population represents the general paramedic population.

Second, because this is a concurrent validity study, 29 C.F.R. § 1607.14(B)(4) requires us to examine whether the test focuses on primary skills learned on the job. In this case, the entrance exam tests three skills: the modified stair-climb, arm endurance, and leg lift. These skills were validated by correlating each skill with all three work samples: the lift and carry, stair-chair push, and stretcher lift. There is a statistically significant correlation between these physical skills and these work samples. In this, the study is fine.

On the issue of reliability, however, the lift and carry poses a problem: its reliability score is only 0.503. That is a 50/50 chance of reliability. Federal regulations direct that reliability, in validity studies like this one, "should be a matter of concern to the user." 29 C.F.R. § 1607.14(5). This is particularly true when the reliability of this lift and carry is equivalent to the proverbial coin toss: heads, it is reliable; tails, it isn't. Further, there was no apparent effort to separate the lift and carry from the rest of the study. Because each of the three skills (modified stair-climb, arm endurance, and leg lift) was validated by correlating it against all three work samples (lift and carry, stair-chair push, and stretcher lift), the unreliability of this lift and carry undermines all three skills that Chicago tests in its physical entrance exam. "All tests must be statistically examined for evidence of reliability before the test developer can establish the validity of the test." *Gillespie v. Wisconsin*, 771 F.2d 1035, 1041 (7th Cir. 1985). Given the lack of reliability in her

study, the test developer in this case cannot establish the validity of her study.

Even if reliability was fully established in this case, validity would be a problem. The plaintiffs legitimately question whether the work samples themselves are a valid measure of job skills. The problem here is that Chicago used the work-sample tests to validate the skills tests—without ever validating the work samples. As a result, we cannot conclude that these work samples reflect "the primary focus of" paramedic skills learned on the job. 29 C.F.R. § 1607.14(B)(4).

Chicago "would have this court find job-relatedness on the basis of a high correlation between the results of two separate testing practices, neither of which by itself has been validated according to accepted methods. We cannot accept this flawed argument." *Guardians Ass'n of N.Y.C. Police Dep't, Inc. v. Civil Serv. Comm'n of the City of New York*, 633 F.2d 232, 244 (2d Cir. 1980). In this case, Chicago created a skills test and a work-sample test, found a strong correlation between the skills test and the work-sample test, and thus concluded that the skills test is a good measure of job-related skills. As the plaintiffs argue, this is a statistical form of self-affirmation. There is no evidence that the work-sample test, which Chicago used to validate the skills test, is a proper validation of job skills. On the contrary, we question whether the work samples actually test the skills that Chicago paramedics learn on the job, as expressly required by the language of Section 1607.14(B)(4).

Gebhardt surveyed Chicago paramedics about on-the-job situations when creating her study. Paramedics indicated that, when getting to patients, they carry equipment between 10 and 100 feet about 73% of the time. If they carry equipment

upstairs, they climb one or two floors about half the time, and climb three or four floors about a third of the time.

Paramedics reported that their patients weighed less than 150 pounds about 24% of the time. Patients weighed 160 to 200 pounds about 40% of the time; 210 to 250 pounds about another 22% of the time; and 260 or more pounds about 14% of the time. Chicago paramedics further reported that, if a patient had to be carried on paramedic equipment, the distance traveled was less than 100 feet about 80% of the time.

Gebhardt also found that, when patients were moved into ambulances, they were in wheeled stretchers about 27% of the time and in stair chairs about 68% of the time. On the record, the use of stair chairs in Chicago appears limited. According to Gebhardt's findings, stair chairs are used for transporting patients into ambulances, usually through the side door. She did not indicate that they were ever used past this point. According to the plaintiffs' evidence at trial, "they don't have ramps at [Chicago] hospitals, and you never push a person into a hospital on a stair chair." *See* Dist. Ct. Docket 554-4 at 665; *see also* Dist. Ct. Docket 557-1 at 226 ("stair chairs are not to be used to transport patients from the ambulance to the hospital under the relevant EMS System Policies and Procedures" and "[p]atients need to be on the stretcher or sometimes secured on the ambulance bench").

Given these actual on-the-job skills, it is difficult to see how all three work samples test job-related skills, much less skills that are a "primary" focus of skills that Chicago paramedics learn on the job. *See* 29 C.F.R. § 1607.14(B)(4). We address the work samples in turn, by comparing the work-sample skills with the skills that paramedics actually use.

To begin with, the lift and carry seems reasonably job- related: it tests paramedics' ability to carry equipment up and down stairs. The plaintiffs object to the timed nature of this test. In the context of disparate-impact litigation, the Eighth Circuit has cast doubt on the value of timed physical-skills testing. As it said, "where hiring is contingent upon test performance, applicants tend to work as fast as possible during the test in order to outperform the competition." *E.E.O.C. v. Dial Corp.*, 469 F.3d 735, 739, 742–43 (8th Cir. 2006). And as Gebhardt's own study report recognized, "physical test cutoff scores should be set at the minimum acceptable level," not at the maximum level, because the goal is to identify people with sufficient physical skills. Further, faster performance is not always the most careful performance. We can see where some speed, though not excessive speed, may be job-related for paramedics who answer time-sensitive emergency calls. But in staying faithful to the record, we do not have the information necessary to analyze and reach a conclusion on the appropriateness of this timed test. Regardless of how appropriate it was to time the lift and carry, this work sample did not prove reliable. And an unreliable assessment tool cannot be validated. *Gillespie*, 771 F.2d at 1041.

Next, the stair-chair work sample focuses on pushing a stair chair through a course and over a ramp. The plaintiffs also object to the timed nature of this test, but again, we lack the information needed to reach a conclusion on timing. We do recognize that stair chairs are, as their name suggests, designed to be carried up and down stairwells. The record indicates that there are no ramps leading into Chicago hospitals. Even if there were ramps, Chicago paramedics may not transport patients into hospitals using stair chairs. The record does not indicate that the stair chairs in the work sample were

transported for the same distances as stair chairs in real paramedic situations, which casts doubt on this work sample, too. We do not conclude, however, that the stair-chair work sample is necessarily unrelated to on-the-job skills.

Moving to the final work sample, we must conclude that the stretcher lift does not resemble skills learned on the job. Real paramedics raise a stretcher and then move. We question why paramedics in the work sample have their arms "locked." Regardless, when paramedics transport a real patient, they do not cycle the patient-laden stretcher up and down, and the record shows that they typically travel less than 100 feet. It is hard to imagine paramedics requiring nearly four-and-a-half minutes to cross 100 feet. Yet the record indicates that paramedics in the stretcher-lift sample had a different task: they had to complete 13 cycles up and down, which requires a total of 4 minutes and 20 seconds to complete, even with rest times omitted.

Further, stretchers are usually carried by at least two paramedics and/or wheeled, while the stretcher-lift work sample requires one paramedic to carry all the weight alone. At the beginning of the stretcher-lift work sample, paramedics are lifting 90 pounds alone. This is the equivalent of carrying a 180-pound patient in tandem. By the end of this lift work sample, the paramedic is lifting 220 pounds alone. This is the equivalent of carrying a 440-pound patient in tandem, immediately after carrying 12 other patients who were increasingly heavy. There may be situations when paramedics transport more than a dozen patients in rapid succession, where most of the patients are atypically heavy individuals, but this is not within the scope of "primary" EMS skills. *See* 29 C.F.R.

§ 607.14(B)(4). For this work sample, the simulated job skills and real job skills are different in description.

In comparing the skills that Chicago paramedics learn on the job with the skills that Gebhardt's three work samples require, we must conclude that these are two different sets of skills. Even if they were the same, the work-sample skills are more taxing than real on-the-job skills. Gebhardt has test subjects cycling abnormally heavy stretchers up and down for 4 minutes and 20 seconds, for example, even with rest times omitted. In contrast, paramedics usually transport relatively lighter stretchers across a distance of 100 feet or less, which should require substantially less than 4 minutes. And as one of our sister circuits has affirmed, plaintiffs should prevail on their disparate-treatment claim when a physical-skills entrance exam is "significantly more difficult than the actual job workers performed at the plant." *Dial*, 469 F.3d at 739, 742–43. The difficulty of these work samples also undermines Chicago's argument that these work samples represent real skills that Chicago paramedics learn on the job.

In this case, at least two out of three work samples are not valid. The validity of the three skills that are tested in Chicago's entrance examination, however, depends on all three work samples being valid. This undermines the entire physical-skills entrance test that Chicago administers.

The physical entrance exam that resulted from this study of volunteer paramedics risks cementing unfairness into Chicago's job-application process. Unfairness is defined this way: when women characteristically obtain lower scores on the physical entrance exam than men, and the score differences "are not reflected in differences in a measure of job performance," the entrance exam is unfair. *See* 29 C.F.R.

§ 1607.14(B)(8)(a). We recognize that, in itself, there is nothing unfair about women characteristically obtaining lower physical-skills scores than men. But the law clearly requires that this difference in score must correlate with a difference in job performance.

To guard against this unfairness, the law requires that the physical exam must validly test job-related skills. We recognize that, if men and women have adequate physical skills together, patients can benefit from coed paramedic teams. The minimum requirement is adequacy, not superiority. *See Lanning v. Se. Pa. Transp. Auth.*, 308 F.3d 286, 287 (3d Cir. 2002) (affirming that, in a disparate-impact claim, "a discriminatory cutoff score on an entry level employment examination must be shown to measure the minimum qualifications necessary for successful performance of the job"). Perhaps mixed-gender teams could offer patients a more diverse combination of physical and psychological care than single-gender teams. A female paramedic might fit into a space where a male paramedic does not; a female victim might be helped by having a female paramedic on the team. And in this case, the validated testing of job-related skills simply is not there: it is not enough to show a strong correlation between two tests that Chicago created concurrently. To validate the other test, at least one test must itself be a valid measure of job skills.

Accordingly, on this detailed review of the record, we conclude that there is clear error in the factual conclusions reached below. Under the federal requirements for validity studies articulated in 29 C.F.R. § 1607.14(B)(4), clear problems arise with the job analysis, the reliability of this validity study, and the validation of Chicago's validity study. Chicago failed

to establish that its physical-skills entrance test reflects "'important elements of job performance.'" *Dial*, 469 F.3d at 743 (quoting 29 C.F.R. § 607.5(B)). And this lack of connection between real job skills and tested job skills is, in the end, fatal to Chicago's case. Thus, the plaintiffs should have prevailed on their Title VII disparate-impact claims.[7]

C. Disparate Treatment and Impact: Evidentiary Rulings

Finally, the defendants object to evidentiary rulings made below. We address these briefly, bearing in mind that the standard of review for evidentiary rulings is abuse of discretion, and we affirm the evidentiary rulings appealed here. *See Bradley v. Work*, 154 F.3d 704, 708–09 (7th Cir. 1998).

First, during the disparate-treatment trial, the district court admitted handwritten committee notes under the business-records exception. These notes were written by Deputy Fire Commissioner Derrick Jackson during a Chicago Fire Department committee discussion about physical injuries that Chicago's paramedics sustained on the job. First Deputy Commissioner Charles Stewart authenticated the handwriting as Jackson's. He further testified that, as part of its regular business practices, the Chicago Fire Department maintains committee-meeting notes like Jackson's notes.

---

[7] The plaintiffs also challenge the passing score established for this skills test, by arguing that there is no appropriate statistical basis for setting the cut-off score. Because this particular test is not validated, however, we do not address the statistics behind the passing score.

Chicago did not carry its burden of proof on the test-validation issue, so we also do not reach the parties' arguments regarding less-discriminatory alternatives to the test.

Meeting minutes properly fall within the business-records exception. *United States v. Borrasi*, 639 F.3d 774, 779 (7th Cir. 2011). Federal Rule of Evidence 803(6) requires, among other things, that these notes must be made during or near the time of the meeting by someone with knowledge. The notes must also be maintained in the regular course of business. The disputed evidence satisfies Rule 803's requirements. The plaintiffs rely on *United States v. Borrasi*, however, where we excluded evidence of a committee report that was discussed in committee minutes. *See id.* at 779–80.

In *Borrasi*, we concluded that "[t]hose reports and any statements therefrom are hearsay, as each comprises statements written by [individuals] not testifying before the court that [a party] wished to introduce for the truth of the matters asserted." *Id. Borrasi* is distinguishable. There, the moving party sought to admit a report through the committee notes. Here, Chicago sought to admit evidence of what was discussed in the committee meeting itself. This was proper.

Though the plaintiffs object to Chicago admitting the evidence through Stewart, saying that Stewart never attended the meeting in question and could not be cross-examined about the meeting, Stewart's role was simply to authenticate the business-record evidence. If the plaintiffs wished to cross-examine Jackson, they could have subpoenaed him as a witness. The district court did not abuse its discretion.

Second, during the disparate-treatment trial, the plaintiffs sought to admit evidence that Gebhardt had previously engaged in conduct that reduced the number of jobs for which women qualified. The district court would not let them offer evidence that Gebhardt's work adversely affected women. It

allowed Gebhardt to provide this compelling testimony, however, regarding the purpose of her entire career:

> Question: What is the purpose in your mind of your entire career with respect to developing and validating physical abilities tests for physically demanding jobs?
>
> Answer: Well, it's to actually give women a shot at a physically demanding job. And it occurred when I first started doing this, I was doing a project for AT&T. And women wanted to work in what's called the outside craft positions. They were working clerical, and the pay was a lot better for outside craft….
>
> One of the things I saw was that the women could do the jobs…. And I also saw in my early years that employers said: Oh, well we want to get women into these jobs.
>
> So what happens is they just put them in the jobs…. the men were like: Oh, it's nice you're here.
>
> And then they became resentful of the fact that some of the women—not all, but some—could not handle the physical demands….
>
> So what we had was when we started putting in the physical abilities tests, we found that one, the women could meet the physical demands and they were much more accepted by their peer—male peers when they could actually perform the tasks. And that was a good thing because in the early years, they got a bad attitude towards it.

> So to suggest that I would discriminate against women is absurd because I've spent most of my life trying to make sure that women are successful in these jobs.

Dist. Ct. Docket 554-9 at 26–27. With the defense offering this evidence about the purpose of Gebhardt's "entire career," the plaintiffs should also be allowed to offer evidence that rebuts her testimony. Only permitting evidence on one side of this equation would be unfairly prejudicial. The plaintiffs do not provide specific explanations, however, of what they tried to admit into evidence and why the district court ruled against them. As a result, the objection is undeveloped and is thus waived for purposes of the first jury trial.

Third, the plaintiffs sought to admit evidence that the Chicago Fire Department offers more pretest training to its firefighter applicants than its paramedic applicants. Once again, however, the plaintiffs offer a generalized objection. They do not claim that the two tests are comparable. Nor do they explain why the Chicago Fire Department should approach two tests in the same way. On the face of the matter, it is reasonable for an employer to have different hiring practices for different positions. Again, this objection is undeveloped, and we find that it is waived for the first jury trial.

The remaining objections appear to be against evidentiary rulings made during the bench trial on disparate impact. Because we conclude that the concurrent validity study was not validated, we do not address these rulings.

Conclusion

The disparate-treatment claims are REMANDED for a new trial, with directions to read the original version of Jury Instruction 24. The disparate-impact trial verdict is REVERSED, with instructions to enter judgment in the plaintiffs' favor. Finally, the evidentiary rulings below are AFFIRMED.