UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK
----------X
GULINO, ET AL.,

Plaintiffs,

96-CV-8414 (KMW)
OPINION & ORDER

-against-

THE BOARD OF EDUCATION OF THE CITY SCHOOL DISTRICT OF THE CITY OF NEW YORK,

Defendant.
----------X
WOOD, U.S.D.J.:

From 1993 to 2012, New York City's Board of Education (the "BOE") required all applicants for public school teaching positions to pass a qualifying examination called the Liberal Arts and Sciences Test, often referred to as the "LAST." There were two incarnations of the exam: the LAST-1, administered from 1993–2004, and the LAST-2, a significantly revised version administered from 2004–2012. These tests were not intended to evaluate an applicant's mastery of the particular subject areas she might teach, or an applicant's capacity to respond to pedagogical challenges that might arise in the classroom—the BOE evaluated those abilities with separate qualifying examinations. Rather, as their full names suggest, the LAST-1 and LAST-2 were designed solely to test an applicant's understanding of the liberal arts and sciences.

Judge Motley of this court previously held that the BOE unfairly discriminated against African-American and Latino applicants, in violation of Title VII of the Civil Rights Act, by requiring them to pass the LAST-1.[1] *Gulino v. Bd. of Educ. of City Sch. Dist. of City of New*

[1] This case was originally tried before Judge Motley. Judge Motley passed away in 2005. After appeal, the case was assigned to Judge Stein on remand, who later transferred it to the undersigned in 2009.

York, 907 F. Supp. 2d 492, 498 (S.D.N.Y. 2012) (Wood, J.) ("*Gulino III*"), *aff'd sub nom. Gulino v. Bd. of Educ. of New York City Sch. Dist. of City of New York*, 555 F. App'x 37 (2d Cir. 2014). Under Title VII, a plaintiff may make a *prima facie* showing of discrimination by demonstrating that a qualifying examination has a disparate impact on minority applicants. Plaintiffs made such a *prima facie* showing at trial in 2003 (before Judge Motley) by proving that African-American and Latino test takers passed the LAST-1 at significantly lower rates than other groups. Title VII permits a defendant, in turn, to defend against a *prima facie* showing of discrimination by proving that a qualifying examination was properly validated as job related—in other words, that the exam's designers used adequate procedures to ensure that it would test only the knowledge, skills, and abilities necessary for competent job performance. The BOE failed to defend the LAST-1 in this way at trial. Although some familiarity with the liberal arts and sciences is no doubt valuable for many teachers, the BOE did not demonstrate that the LAST-1's designers had employed procedures to identify the *specific* areas of the liberal arts and sciences that *any* competent teacher, regardless of grade level or subject area, would need to understand. Accordingly, in 2012, the Court held that Plaintiffs had prevailed under Title VII.

Exercising its broad remedial authority, the Court then appointed a neutral expert, Dr. James Outtz, who was acceptable to the parties, to evaluate whether the LAST-2 also had a disparate impact on African-American or Latino test takers — and if so, whether the exam had been properly validated as job related. The Court permitted the BOE to submit a rebuttal expert report from Dr. Chad Buckendahl, and held a hearing during which both parties and the Court questioned the experts. Dr. Outtz concluded that the LAST-2 had a disparate impact on African-American and Latino test takers and had not been properly validated as job related. Dr.

Buckendahl and the BOE did not dispute the exam's disparate impact, but they argued that the LAST-2 had been properly validated.

After reviewing all of the evidence offered by Dr. Outtz and the parties, including expert opinions and the Equal Employment Opportunity Commission's Uniform Guidelines on Employee Selection Procedures, the Court holds that the BOE unfairly discriminated against African-American and Latino applicants by requiring them to pass the LAST-2. Like its predecessor, the LAST-2 had a disparate impact on African-American and Latino test takers. And like its predecessor, the LAST-2 was not properly validated as job related, because the exam's designers did not employ procedures to identify the *specific* areas and depth of knowledge of the liberal arts and sciences that any competent teacher would need to understand. The BOE's use of the LAST-2 was thus unfairly discriminatory under Title VII.

In reaching that conclusion, the Court does not suggest that it would be unhelpful or unwise for the BOE to test applicants' knowledge of the liberal arts and sciences with a properly validated exam. It may be the case that all teachers, whether they instruct kindergarteners or high school seniors, must understand certain areas of the liberal arts and sciences (separate and apart from the particular subject matter they teach) in order to be competent in the classroom. But the Court is not permitted to simply intuit that fact; test designers must establish it through adequate validation procedures. In that regard, both the LAST-1 and the LAST-2 were deficient, which renders them indefensible under Title VII.

I. NEW YORK STATE'S TEACHER LICENSURE EXAMINATIONS[2]

The New York State Education Department ("the SED") requires the BOE to hire only New York City public school teachers who have been certified by the State. *Gulino III*, 907 F. Supp. 2d at 498. If the BOE were to hire teachers who have not been certified by the State, New York City could lose as much as $7.5 billion a year in state funding. *See* (Oct. 23, 2014 Jt. Ltr. [ECF No. 515] at 2–3).

Beginning in 1993, the SED required teachers seeking certification to pass the LAST-1, a new test developed at the State's request by National Evaluation Systems ("NES"),[3] a professional test development company. *Id.* at 499–500. The LAST-1 "include[d] questions related to scientific, mathematical, and technological processes; historical and social scientific awareness; artistic expression and the humanities; communication and research skills; and written analysis and expression." (Foley Decl., Ex. I ("Clayton Decl.") [ECF No. 377-3] at ¶ 4).

In 2004, the SED phased out the LAST-1 and replaced it with the LAST-2. *See* (Dec. 8, 2009 Order [ECF No. 243] at 3). The LAST-2 was first used for teacher certification on February 14, 2004. (*Id.*) Prior to using the LAST-2, NES and SED documented the process by which they sought to validate the test as job related. *See generally* (Clayton Decl.).

At the time the LAST-2 was implemented, prospective teachers were required to pass two additional written exams: the Assessment of Teaching Skills – Written ("ATS-W"), and the Content Specialty Test ("CST") applicable to the teacher's subject area. *See* (BOE Ltr., Attachment A, [ECF No. 504-1]) (listing the different certification requirements mandated by the

---

[2] For a more detailed discussion of the history of teacher licensure requirements in New York state, see *Gulino III*, 907 F. Supp. 2d at 498–500.

[3] NES was acquired by NCS Pearson, Inc. ("Pearson") in April 2006. *See Pearson Enters Teacher Certification Market by Acquiring National Evaluation Systems*, www.pearson.com (April 25, 2006), https://www.pearson.com/news/announcements/2006/april/pearson-enters-teacher-certification-market-by-acquiring-national.html.

SED over time). According to Pearson, the ATS-W was "designed to assess pedagogical (teaching) skills that New York educators determined to be important to the adequate performance of the job of . . . public school teachers." (Pearson Ltr. [ECF No. 500] at 2). The CST was designed to "assess the specific knowledge and skills needed to teach specific subject matter in New York public schools, such as mathematics, physics, chemistry, American Sign Language, Cantonese, Japanese, etc." (*Id.*) A prospective teacher was required to pass the ATS-W, any applicable CST, and the LAST-2 in order to receive a teaching license. Applicants were not permitted to compensate for a poor score on one exam with a high score on another. *See* (Feb. 3, 2015 Ltr., Attach. I ("Outtz Report") [ECF No. 549-1] at 37).

II. PROCEDURAL HISTORY

The nineteen-year history of this case is long and winding, and has been set out in the Court's prior opinions, familiarity with which is assumed.[4] What follows is a condensed recounting of that history, as it relates to the current issues at bar.

A. The LAST-1

Plaintiffs, who represent a class of African–American and Latino applicants for teaching positions in the New York City public school system, originally alleged that the BOE had violated Title VII by requiring applicants to pass the LAST-1. Plaintiffs claimed that the exam had a disparate impact on African-American and Latino test takers, which was unfairly discriminatory because the exam was not job related.[5]

---

[4] *See Gulino III*, 907 F. Supp. 2d 492 (finding BOE liable under Title VII on remand); *Gulino II*, 460 F.3d 361 (partially affirming and reversing Judge Motley's original liability decision); *Gulino v. Bd. of Educ. of the City Sch. Dist. of the City of N.Y.*, No. 96-CV-8414, 2003 WL 25764041 (S.D.N.Y. Sept. 4, 2003) (Motley, J.) ("*Gulino I*") (original liability opinion).

[5] The Second Circuit, in a previous decision in this case, held that the BOE could be found liable for New York State's requirement because "Title VII preempts any state laws in conflict with it"; "even though BOE was merely following the mandates of state law, in using the LAST to certify teachers, it was nevertheless subject to

The case was initially assigned to the Honorable Constance Baker Motley in 1996. In 2003, following "an epic bench trial that lasted more than eight weeks and filled over 3,600 pages of trial transcript," *Gulino I*, 2003 WL 25764041, at *1, Judge Motley ruled that the BOE had not violated Title VII by adopting the SED's requirement that teachers pass the LAST-1[6] in order to receive a permanent license.[7] *Id.* at *30–31 ¶¶ 161–64. Although Judge Motley held that Plaintiffs had established a *prima facie* case of disparate impact, *id.* at *30 ¶ 160, she ultimately found that the LAST-1 was not unfairly discriminatory because it qualified as job related. *Id.* at *30–31 ¶¶ 161–63.

On appeal, the Second Circuit affirmed in part and reversed in part. Relevant to the instant proceedings, the panel held that Judge Motley had erred by not assessing the LAST-1's job-relatedness under the standard established in *Guardians Association of New York City Police Department, Inc. v. Civil Service Commission of the City of New York* ("*Guardians*"), 630 F.2d 79 (2d Cir. 1980), and remanded so that the district court could apply that standard. *Gulino II*, 460 F.3d at 385, 388.

On remand, this Court held that the LAST-1 was not job related because it had not been properly validated by the State and NES. Accordingly, the Court concluded the BOE had violated Title VII by requiring prospective teachers to pass the test. *Gulino III*, 907 F. Supp. 2d at 516–23.

Title VII liability." *Gulino v. N.Y. State Educ. Dep't* ("*Gulino II*"), 460 F.3d 361, 380 (2d Cir. 2006) (internal quotation marks and citations omitted).

[6] The LAST-2 had not gone into effect at the time of Judge Motley's ruling, and thus the Court at that time described the LAST-1 simply as "the LAST." The Second Circuit, and this Court on remand, followed Judge Motley's lead. The Court makes the distinction here for clarity's sake.

[7] Plaintiffs initially sued the SED in addition to the BOE, and therefore, Judge Motley's holdings in *Gulino I* applied to both the SED and the BOE. *See generally Gulino I*, 2003 WL 25764041. However, the SED has since been dismissed from this case. *See Gulino II*, 460 F.3d at 388. Accordingly, this Part will focus solely on the procedural history of this case as it relates to the BOE, not the SED.

B. The LAST-2

By the time the Court decided Plaintiffs' challenge to the LAST-1, the SED had retired the exam in favor of the LAST-2. Exercising its remedial authority to require that a "subsequent exam" comply with Title VII, *Guardians*, 630 F.2d at 109, the Court then sought to ensure that the LAST-2 was not unfairly discriminatory. The Court appointed Dr. Outtz to serve as a neutral expert and assess whether the LAST-2 had a disparate impact on African-American or Latino test takers—and if so, whether the exam qualified as job related. *See* (Apr. 29, 2014 Hearing Tr. [ECF No. 428] at 55); (Oct. 29, 2013 Hearing Tr. [ECF No. 403] at 4–8).

On February 3, 2015, Dr. Outtz concluded that the LAST-2 had a disparate impact on African-American and Latino test takers and did not qualify as job related, because it had not been properly validated. *See generally* (Outtz Report). In response, the BOE submitted the report of Dr. Buckendahl, which did not address the issue of disparate impact but argued that the LAST-2 had been properly validated. *See generally* (Buckendahl Response [ECF No. 592]). The SED also submitted a response, which asserted that Dr. Outtz's report was flawed and the LAST-2 had been properly validated.[8] *See generally* (SED Response [ECF No. 589]). The Court held a hearing on March 20, 2015, where both the Court and the parties questioned Dr. Outtz and Dr. Buckendahl about their opinions concerning the validity of the LAST-2.

[8] All of these reports are discussed in greater detail in Part V, *supra*.

III. LEGAL STANDARD

A. Title VII's Burden Shifting Framework

Under Title VII, a plaintiff can make out a *prima facie* case of discrimination with respect to an employment exam by showing that the exam has a disparate impact on minority candidates. *See N.A.A.C.P., Inc. v. Town of E. Haven*, 70 F.3d 219, 225 (2d Cir. 1995).

The defendant can rebut that *prima facie* showing by demonstrating that the exam is job related. *Id.* To do so, the defendant must prove that the exam has been validated properly. Validation requires showing, "by professionally acceptable methods, [that the exam 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.'" *Gulino II*, 460 F.3d at 383 (quoting *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 431 (1975)).

In *Guardians*, the Second Circuit devised a five-part test for determining whether an employment exam, such as the LAST-2, has been properly validated and is thus job related for the purposes of Title VII:

(1) "the test-makers must have conducted a suitable job analysis";

(2) the test-makers "must have used reasonable competence in constructing the test";

(3) "the content of the test must be related to the content of the job";

(4) "the content of the test must be representative of the content of the job"; and

(5) there must be "a scoring system that usefully selects" those applicants "who can better perform the job."

Guardians, 630 F.2d at 95; *see also Gulino II*, 460 F.3d at 384. The first two elements of this test, which concern the quality of the test's development, are "particularly crucial" because "validity is determined by a set of operations, and one evaluates . . . validity by the thoroughness

and care with which these operations have been conducted." *Id.* at 95 n.14 (internal citation and quotation omitted).

Because validation requires expertise that courts lack, validation is "not primarily a legal subject." *Guardians*, 630 F.2d at 89. Accordingly, to determine whether an employment exam is properly validated, a court "must take into account the expertise of test validation professionals." *Gulino II*, 460 F.3d at 383. There are two primary sources of expertise on which courts rely to assess validation: (1) the testimony of experts in the field of test validation; and (2) the Equal Employment Opportunity Commission's Uniform Guidelines on Employee Selection Procedures ("Guidelines"), 29 C.F.R. § 1607,[9] which establish standards for properly validating an employment test. *Gulino III*, 907 F. Supp. 2d at 515. Although courts are not bound by the Guidelines, the Supreme Court has stated that the Guidelines are "entitled to great deference" because they represent "the administrative interpretation of [Title VII] by the enforcing agency." *Griggs v. Duke Power Co.*, 401 U.S. 424, 433–34 (1971). According to the Second Circuit, the Guidelines should be the "primary yardstick by which [to] measure [D]efendant['s] attempt to validate the LAST[-2]" because of their longstanding use in the field. *Gulino II*, 460 F.3d at 384.

B. The Content-Construct Validity Continuum

"The threshold task in determining the validity of a challenged examination is to select the appropriate method for assessing its job-relatedness." *Guardians*, 630 F.2d at 91. The Guidelines detail two validation techniques that are relevant here: content validation and construct validation.

[9] For clarity's sake, the Court will denote sections of the Guidelines as "Guidelines § 1607.X," rather than as "29 C.F.R. § 1607.X."

Guardians defined "content validation" as "a technique appropriate for tests that measure 'knowledges, skills, or abilities' [or 'KSAs'] representative of the 'content' of the job." *Guardians*, 630 F.2d at 92 (quoting Guidelines § 1607.14(C)(1)). It defined "construct validation" as a technique that "attempts to measure 'constructs,' that is, inferences about mental processes or traits, such as 'intelligence, aptitude, personality, commonsense, judgment, leadership and spatial ability.'" *Id.* at 92 (quoting Guidelines § 1607.14(C)(1)). These definitions suggest that content validation is appropriate when testing for job-specific abilities—for example, the ability of a carpenter to measure the dimensions of building materials—and construct validation is appropriate when testing more general, non–job specific abilities, such as spatial reasoning, or problem solving. *See id.* at 92–93 (discussing the relationship between job specificity and the construct-content distinction).

The Second Circuit has accepted the proposition that these two validation methods differ from one another. Content validation requires a test's proponent to show that "the content" of the test "is representative of important aspects of performance on the job for which the candidates are to be evaluated." Guidelines § 1607.5(B). As will be discussed more thoroughly below, this can be demonstrated in a fairly straightforward manner by showing a link between the abilities being tested for and the tasks required by the job in question. *See infra* Part V.c.ii. Construct validity, however, "requires 'an extensive and arduous effort.'" *Guardians*, 630 F.2d at 92 (quoting Guidelines § 1607.14(D)(1)). It demands "a demonstration from empirical data that the test successfully predicts job performance." *Id.* "Developing such data is difficult, and tests for which it is required have frequently been declared invalid." *Id.* (citing cases). The result is that "[t]his content-construct distinction . . . frequently determines who wins the lawsuit. Content validation is generally feasible while construct validation is frequently impossible." *Id.*

Notwithstanding the "sharp distinction" that the Guidelines draw "between tests that measure 'content' . . . and tests that purport to measure 'constructs,'" *Gulino II*, 460 F.3d at 384, *Guardians* concluded that the Guidelines "adopt too rigid an approach." *Guardians*, 639 F.2d at 92.[10] The court noted that "content" and "constructs" "are simply different segments along a continuum reflecting a person's capacity to perform various categories of tasks." *Id.* at 93. On one side of the continuum are those KSAs that are so general, they are relevant to every job. The most common example would be general intelligence. On the other side of the continuum are those KSAs that are so job-specific that they apply to only one job. For instance, only a major league baseball player would need to be able to hit a ninety-five mile-per-hour fastball. Most KSAs, of course, would fall somewhere between those two extremes.

Guardians held that construct validation is not necessary in every instance where a job involves abilities that are somewhat general. "[I]f the test attempts to measure general qualities such as intelligence or commonsense, which are no more relevant to the job in question than to any other job, then insistence on the rigorous standards of construct validation is needed."[11] *Id.* "But as long as the abilities that the test attempts to measure . . . are the most observable abilities of significance to the particular job in question, content validation should be available." *Id.*[12]

---

[10] The Court notes that the distinction seems poorly drawn, in that both types of testing measure abilities and skills.

[11] The court noted that construct validation is necessary in such instances because "tests of this kind are often biased in favor of a person's familiarity with the dominant culture." *Id.* Thus, "permitting them to be used without a showing of predictive validity would perpetuate the effects of prior discrimination." *Id.*

[12] Looked at from a slightly different perspective, it appears that the *Guardians* court was attempting to prevent what it believed to be a harsh result—that simply because an exam tested for somewhat general abilities, that test would almost invariably be found invalid based on a requirement that the test clear the high hurdles construct validation demands. To alleviate some of this harshness, the court attempted to shoehorn general—but not *too* general—abilities into the category of abilities for which content validation is appropriate, even though those abilities might be more appropriately thought of as "constructs," at least as constructs are defined by the Guidelines. Conceiving of content validation in this broader way allows most employment exams to at least stand a chance of being found valid, even if they test for fairly general abilities. *See Guardians*, F.2d at 92 (noting that strict compliance with the Guideline's content-construct distinction would be "inconsistent with Title VII's endorsement of professionally developed tests"). Accordingly, for the purposes of determining whether content validation or construct validation should be required, the simplest, and most straightforward way of interpreting *Guardians* may

Guardians also held that "the degree to which content validation must be demonstrated should increase as the abilities tested for become more abstract."[13] *Id.*[14] In other words, *Guardians* adopted a sliding scale approach to assessing the validity of employment exams; the more general the abilities tested by the exam—which is to say, the less the abilities tested relate only to a particular job—the more rigorously a court should investigate the exam's validity. At a certain point, when the abilities tested become so general that they apply to most any job, content validation is no longer adequate, and construct validation must be used.

IV. THE DEVELOPMENT OF THE LAST-2[15]

A. The Decision to Test Liberal Arts and Science Knowledge

In 1988, a New York State task force studying teacher qualifications determined that all teachers should have a basic understanding of the liberal arts in order to be competent to teach. Commissioner's Task Force on the Teaching Profession, *The New York Report: A Blueprint for Learning and Teaching* ("*Blueprint Report*") 15–19 (1988). It recommended that the state require teachers to pass a liberal arts exam before they receive certification. *Id.* In 1990, the

---

be to acknowledge that almost every employment exam should be assessed based on a content validation methodology. It is only a rare exam that tests for abilities so general that they apply to nearly every job, and thus obligate test proponents to meet the onerous requirements of construct validation.

[13] *Guardians* used the term "abstract," rather than "general," in this particular sentence. However, the decision vacillates between the two terms. *Compare Guardians*, 630 F.2d at 92 ("[I]f the attributes the test attempts to measure are too general, they are likely to be regarded as constructs . . . ."), *and id.* at 93 ("[I]f the test attempts to measure general qualities . . . then insistence on the rigorous standards of construct validation is needed."), *with id.* ("[A]s long as the abilities that the test attempts to measure are no more abstract than necessary, . . . content validation should be available."), *and* ("If the job in question involves primarily abilities that are somewhat abstract, content validation should not be rejected simply because these abilities could be categorized as constructs."). The Court finds the terms "general" or "generalized" to be more precise and accurate than "abstract," and thus, for clarity's sake, the Court will use "general" to characterize those abilities that fall closer to the construct side of the continuum.

[14] *Guardians* noted that this approach was necessary "[t]o lessen the risks of perpetuating cultural disadvantages." *Id.*

[15] Most of the discussion of the development of the LAST-2 is drawn from Dr. Outtz's expert report, and the declaration of Jeanne Clayton, a Senior Area Director for Pearson, who had "overall responsibility for the day-to-day management of NES's involvement with the New York State Education Department." (Clayton Decl. ¶ 1). The Court finds both of these accounts credible and therefore treats their contents, to the extent recounted herein, as established fact.

SED sought to implement the task force's recommendation by contracting with NES to develop the LAST-1. *Gulino III*, 907 F. Supp. 2d at 512–13. Beginning in 1993, the LAST-1 was administered to prospective teachers as a part of their licensure requirement. *Id.* at 499–500. Several years after the LAST-1 was first administered, the Board of Regents of the State of New York issued new regulations governing teacher certification that required the SED to update the exam. (Clayton Decl. ¶ 13). Accordingly, the SED worked with NES to "extensive[ly] redevelop[]" the test through a "process similar in scope" to the initial development of the LAST-1. (*Id.*) That redevelopment, which took place between 2000 and 2004, ultimately resulted in the LAST-2, which was first administered to prospective teachers on February 14, 2004. (Dec. 08, 2009 Order 3).

B. The Development of the LAST-2's Test Framework

To begin developing the LAST-2, the staff at NES first created a test "framework." The staff relied on this framework to construct each of individual test questions contained in the LAST-2. (Clayton Decl. ¶ 10). Ms. Clayton defines a test "framework" as "a document that describes the overall structure and content of a test." (*Id.* ¶ 5). She testified that that framework is then broken down into "major groups of content" called "subareas." (*Id.*) The LAST-2 covered five subareas: (1) "Scientific, Mathematical and Technical Processes;" (2) "Historical and Social Scientific Awareness;" (3) Artistic Expression and the Humanities;" (4) "Communication and Research Skills;" and (5) "Written Analysis and Expression." (Outtz Report 30).

Within each of these five subareas, NES developed a number of "objectives." Ms. Clayton defined "objectives" as "broad statements of the test content that examinees will be tested on." (*Id.* ¶ 6). The objectives for the LAST-2 were "designed to be broad descriptions of

elements of the knowledge and skills that have been determined by New York State educators to be important to the job of a public school teacher in the State of New York." (*Id.*) For example, within the "Scientific, Mathematical and Technical Processes" subarea, one of the objectives states: "Use mathematical reasoning in problem-solving situations to arrive at logical conclusions and to analyze the problem-solving process." (Outtz Report, App. 2, at 51).

Each objective is further delineated in several "focus statements," which "provide details about the nature and range of content covered by the objectives. They are intended to suggest the types of content that may be included in the test items." (Clayton Decl. ¶ 7). Focus statements for the above objective included: "analyzing problem solutions for logical flaws," and "examining problems to determine missing information needed to solve them." (Outtz Report, App. 2, at 51).

NES developed this framework by using two sources. The first was the LAST-1 framework, which NES "thorough[ly] review[ed]" and then "revised." (Clayton Decl. ¶ 16). The second was a set of documents describing common liberal arts and science course requirements at New York state colleges and universities. (Outtz Report 29); (Clayton Decl. 14).[16]

C. Review of the LAST-2's Test Framework

Once the LAST-2's framework was completed, it was reviewed by two committees of New York state educators: the Bias Review Committee ("BRC") and the Content Advisory

[16] Ms. Clayton also claims that NES "consulted numerous materials that define and describe the job of a New York State teacher, including New York State regulations and guidelines for teachers, student learning standards, textbooks and other curricular materials." (Clayton Decl. ¶ 15). The Court has seen no evidence of this beyond Ms. Clayton's statement. Dr. Outtz made no mention of it in his report, and neither Pearson nor SED testified to any of these matters at the hearing on the validity of the LAST-2. Even assuming Clayton's statement is accurate, however, relying on such documents is still a far cry from the job analysis required by *Guardians* and the Guidelines. *See infra* Part V.C (discussing the requirement that job tasks must be ascertained as a part of a proper job analysis).

Committee ("CAC"). (Clayton Decl. ¶¶ 19–24). The BRC evaluated the framework for potential sources of bias, including offensive language, stereotypes, fairness, and diversity. (Outtz Report 31); (Clayton Decl. ¶ 20–21).[17] The CAC, which included educators who "represented subject matter specialties that were broadly diversified," (Clayton Decl. ¶ 19), reviewed the framework for content, accuracy, and appropriateness. (Outtz Report 32).

Next, NES sent out two separate surveys to educators across the state of New York. The goal of these surveys was to "determine from a broader population the importance of the content objectives of [the LAST-2] to the job of a public school teacher in the State of New York." (Clayton Decl. ¶ 26). Each survey asked the respondent to rate the importance of the individual objectives that made up NES's LAST-2 framework. (Outtz Report 32–33). Specifically, the respondent was asked: "How important is the knowledge or skill described by this objective for performing the job of an educator in New York State public schools?" Respondents were asked to rate each objective on a five-point scale, ranging from "no importance" to "very great importance." (Clayton Decl. ¶ 27). Respondents were also asked a more general question: "How well does the set of objectives, as a whole, represent important aspects of liberal arts and sciences knowledge and skills required for performing the job of an educator in New York State public schools?" This, too, was rated on a five-point scale. (*Id.* at 28).

The samples for both surveys, however, were small. The first survey was sent to 500 certified public school teachers, 320 of which (64%) were completed and returned to NES. (Outtz Report 33). *Contrast M.O.C.H.A. Soc'y Inc. v. City of Buffalo* ("*M.O.C.H.A. Soc'y II*"), 689 F.3d 263, 269–70 (2d Cir. 2012) (describing a job analysis survey of firefighters, which was

[17] Dr. Outtz notes in his report that "[n]o information was provided as to why the persons chosen for the BRC were considered to have expertise in making the judgments they were requested to make." (Outtz Report 31–32).

sent out to 5,934 individuals, and completed and returned by 2,502 individuals). Only twenty-four of the respondents were African-American, and only ten were Latino. (Outtz Report 33). The responses from these two groups were not analyzed separately to determine if their responses differed from those of Caucasian respondents. (*Id.*) The second survey was distributed to 181 faculty members, but only 45 (25%) were returned. (*Id.*) None of the survey responses came from African-American faculty members, and only three came from Latino faculty. (*Id.* at 33–34).

Survey results indicated that respondents believed all of the objectives were are least somewhat important, and most of them were of "great importance." (Outtz Report 34); (Clayton Decl. ¶¶ 30–31).

D. Using the Framework to Develop Test Questions

After the SED approved the framework, NES began the process of item development, whereby the individual test questions were drafted, reviewed, and refined. It appears that some of these questions were derived from test questions in the existing LAST-1 item bank. Ms. Clayton states that LAST-1 questions "were given preliminary designations for continued use, for revision, or for deletion" based on their relevance to the LAST-2 framework. (Clayton Decl. ¶¶ 34–35). The newly-drafted test questions were then reviewed by the BRC and the CAC. (*Id.* ¶ 37). However, it appears that those LAST-1 questions that were designated for continued use were reviewed only by the CAC. *See* (*id.* ¶ 45) ("All items designated for continued use from the existing [LAST-1] item bank were reviewed by the [LAST-2] Content Advisory Committee to verify their continued match to the revised objectives and their continued job-relatedness, accuracy and freedom from bias.").

Next, the new and revised test questions were "pilot tested," in a two-pronged process. NES included some of the potential questions in officially-administered LAST-1 exams, but designated those questions as non-scorable items, such that they did not count towards a test-taker's score. Additional questions were separately administered to volunteer examinees as a means of independently analyzing how test takers responded to the questions. (*Id.* ¶ 46). The results from this pilot testing were reviewed by the BRC and the CAC. (*Id.* ¶¶ 49–50).

Finally, NES created a Passing Score Review Panel. The Panel consisted of New York educators, who provided the information the New York Commissioner of Education used to set the passing score for the LAST-2. (*Id.* ¶ 52). The Panel was asked to "[i]magine a hypothetical individual who is just at the level of knowledge and skills required to perform the job of an educator receiving a teaching certificate in New York State. What is the number of multiple-choice items on the test that would be answered correctly by this individual?" (Outtz Report 39). Dr. Outtz notes that this is a "typical process for setting a passing score." (*Id.* at 41).

V. ANALYSIS

Based on the reports and testimony of Dr. Outtz and Dr. Buckendahl, and the submissions made by both parties and the SED with respect to the development of the LAST-2, the Court finds that Plaintiffs have made a *prima facie* showing of discrimination, by demonstrating that the exam causes a "'disparate impact on the basis of race, color, religion, sex, or national origin.'" *Ricci v. DeStefano*, 557 U.S. 557, 578 (2009) (quoting 42 U.S.C. § 2000e-2(k)(1)(A)(i)). The BOE has failed to rebut that *prima facie* showing because it has not demonstrated that the LAST-2 was properly validated. As explained below, NES's test development process did not comport with the five-factors the *Guardians* court deemed critical to exam validation.

A. Plaintiffs Have Made a *Prima Facie* Showing of Discrimination

A *prima facie* showing of discrimination "requires plaintiffs to establish by a preponderance of the evidence that the employer uses a particular employment practice that causes a disparate impact on the basis of race, color, religion, sex, or national origin." *United States v. City of N.Y.* ("*Vulcan Soc'y*"), 637 F. Supp. 2d 77, 86 (E.D.N.Y. 2009) (internal quotation marks omitted). To do so, a party must "(1) identify a policy or practice, (2) demonstrate that a disparity exists, and (3) establish a causal relationship between the two." *Id.* (internal quotation marks omitted). A party can meet the second and third requirement by relying on the "80% rule." The rule is described in Guidelines § 1607.4(D):

> A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.

In other words, "if the minority group performs less than 80% as well as the highest performing group, disparate impact will generally be inferred." *Vulcan Soc'y*, 637 F. Supp. 2d at 87.

Plaintiffs have satisfied all three of the requirements, as is demonstrated in Dr. Outtz's report. Dr. Outtz's report identifies the specific employment practice at issue: the BOE's requirement (mandated by the SED) that prospective teachers pass the LAST-2 before they can be hired. (Outtz Report 5). Although the pass rates varied year to year, Dr. Outtz's report demonstrates disparate impact by showing that the pass rates for African-American and Latino applicants were between 54% and 75% of the pass rate for Caucasians. (*Id.* at 14–16); (Outtz Rebuttal [ECF No. 565] at 16–22).

The SED disputes Dr. Outtz's calculations and argues that Dr. Outtz should have "take[n] into account the *best* attempt by candidates prior to applying for a license to teach," rather than their *first* attempt, as Dr. Outtz did. (SED Response 1). Dr. Outtz disagrees, stating:

> I consider the first attempt to be the correct metric because use of all attempts discounts the additional time and effort that must be expended after an attempt fails. Additional attempts may also mean falling behind cohorts who pass on an earlier attempt in terms of accruing seniority and eligibility for promotion.

(Outtz Rebuttal 22). The Court agrees with Dr. Outtz's rationale, and holds that it was proper for him to calculate adverse impact based on first attempts. *See Ass'n of Mexican-Am. Educators v. California*, 937 F. Supp. 1397, 1407 (N.D. Cal. 1996), *aff'd*, 231 F.3d 572 (9th Cir. 2000) ("[A]dverse impact is appropriately measured by the first time a candidate sits for the [employment exam] and fails it.").

Accordingly, the Court holds that Plaintiffs have met their *prima facie* burden of demonstrating disparate impact. The burden then shifts to the proponent of the test to demonstrate that the test was "job related for the position in question and consistent with business necessity." 42 U.S.C. § 2000e–2(k)(1)(A)(i); *see also M.O.C.H.A. Soc'y II*, 689 F.3d at 274. The BOE has failed to make such a showing.

B. Defendants Have Failed to Rebut Plaintiffs' *Prima Facie* Showing by Establishing That the LAST-2 is Job Related

To prove job-relatedness, a test proponent must demonstrate the test's compliance with each of the five factors set forth in *Guardians* and discussed in Part III, *supra*. The Court's analysis of the LAST-2 here focuses primarily on the first of those five factors: the sufficiency of NES's job analysis.

NES's job analysis involved the use of a content validation methodology to validate the LAST-2. In light of the content-construct continuum described by *Guardians*, this was the

correct methodology to use. An understanding of the liberal arts and sciences is not a KSA *so* general that it is needed to perform nearly every job.

The SED contends that the LAST-2 tests for specific content, such as "math, science, and technology," "art, literature, religion, and philosophy," and "geography and culture." (SED Ltr. [ECF No. 590] at 5); *but see* (*id.*) ("The LAST measures general knowledge."). After reviewing several LAST-2 exams, the Court finds that although the *texts* as to which an applicant is questioned are on topics such as math, art, etc., *the questions themselves* do not appear to require any significant outside knowledge to answer correctly. They test less for content than for such abilities as reading comprehension, logical thinking, and problem solving. *See* (March 20, 2015 Hearing Tr. 26–27). Those abilities are quite general; many, if not most, jobs require at least some level of reading comprehension, for instance. Based on *Guardian*'s sliding scale approach, the Court therefore must rigorously assess the LAST-2's content validity, beginning with NES's job analysis.

i. NES Did Not Perform a Suitable Job Analysis

A job analysis is an "assessment 'of the important work behavior(s) required for successful performance'" of the job in question and the "'relative importance'" of these behaviors. *Guardians*, 630 F.2d at 95 (quoting Guidelines § 1607.14(C)(2)). The purpose of a job analysis is to ensure that an exam adequately tests for the KSAs that are actually needed to perform the daily tasks of the job. *See Vulcan Soc'y*, 637 F. Supp. 2d at 111. The test developer must be able to explain the relationship between the subject matter being assessed by the exam and the job tasks identified. *Compare id.* (finding that defendant's job analysis for a test given to firefighter candidates was inadequate because no effort had been made to explain the relationship between the knowledge, skills, and abilities being tested on the exam and the tasks involved in

being a firefighter), *with M.O.C.H.A. Soc'y Inc. v. City of Buffalo* ("*M.O.C.H.A. Soc'y I*"), No. 98-CV-99C, 2009 WL 604898, at *14 (W.D.N.Y. Mar. 9, 2009) (finding "comprehensive" job analysis adequate where employer had conducted multiple surveys, statistical analyses, and solicited committee input to ensure subjects evaluated by exam related to the tasks involved in being a firefighter). This requirement ensures that "the pertinent abilities have been selected for measurement." *Vulcan Soc'y*, 637 F. Supp. 2d at 111.

To perform a suitable job analysis, a test developer must: (1) "identify the tasks involved in performing the job," *Gulino III*, 907 F.2d at 516; *see also Guardians*, 630 F.2d at 95 (describing defendant's job analysis as adequate because, *inter alia*, "the work behaviors involved . . . were identified by extensive interviewing, and subjected to serious review"); (2) "includ[e] a thorough survey of the *relative importance* of the various skills involved in the job in question," *M.O.C.H.A. Soc'y, II*, 689 F.3d at 278 (internal quotation omitted) (emphasis added); and (3) define "the *degree of competency* required in regard to each skill." *Id.* (internal quotation omitted) (emphasis added).

NES Did Not Identify Any Job Tasks

As Dr. Outtz points out in his report,[18] the core flaw in NES's job analysis was that it failed to identify *any* job tasks whatsoever. (Outtz Report 29). Without identifying the tasks

[18] The Court gives significant weight to the conclusions of Dr. Outtz's report. As the Court-appointed neutral expert in this case, Dr. Outtz is impartial, and the Court finds him to be reliable. *See In re Ephedra Prods. Liab. Litig.*, 478 F. Supp. 2d 624, 634 (S.D.N.Y. 2007) (Rakoff, J.) (discussing the reliability of a neutral expert's opinion because it was "formed without any monetary inducement"); *United States v. Jones*, 876 F. Supp. 395, 399 (N.D.N.Y. 1995) (noting that, as compared to the defendant's expert, "a court appointed expert is in a better position to render an impartial opinion"); *United States v. Mosley*, 500 F. Supp. 601, 605 (N.D.N.Y. 1980) ("This Court gives the greater weight to the testimony of the Court appointed experts whose job it is to impartially consider all of the . . . evidence and to advise the Court in accordance therewith."); *cf. Kaufman v. Edelstein*, 539 F.2d 811, 818 (2d Cir. 1976) (stating that "[t]he situation of the court appointed expert . . . differs utterly from that of an expert called by a party," and that a court appointed expert is expected to "arrive at an informed and unbiased opinion"); *see also Gulino II*, 460 F.3d at 383 ("Because of the substantive difficulty of test validation, courts must take into account the expertise of test validation professionals.").

involved in performing the job (required by the first factor discussed above), it was not possible for NES to determine the *relative importance* of each job task (second factor), or to define the *degree of competency* required for each skill needed to accomplish those job tasks (third factor). Accordingly, the Court finds NES's job analysis to be wholly deficient.

Instead of beginning with ascertaining the job tasks of New York teachers, the two LAST examinations began with the premise that all New York teachers should be required to demonstrate an understanding of the liberal arts. The impetus for the LAST-2, as it was for the LAST-1, appears to have been the 1988 report by the Commissioner's Task Force on the Teaching Profession, discussed in Part IV, *supra*, which recommended that New York include a liberal arts requirement in its licensing procedure.[19]

In time, the SED adopted the Commission's recommendation, and contracted with NES to design the LAST-1, and then eventually, the LAST-2. NES began developing the LAST-2, as it had for the LAST-1, by consulting documents describing liberal arts and general education undergraduate and graduate course requirements, syllabi, and course outlines. *See* (Clayton Decl. ¶ 14); (Outtz Report 29). NES then defined the KSAs it believed a liberal arts exam should assess, based on the way the liberal arts were characterized in those documents. NES used those KSAs to create the test framework that the BRC and CAC reviewed. That framework was later the subject of the survey NES sent out to educators, inquiring about the importance of the objectives and focus statements contained within the framework. Thus, NES did not investigate the job tasks that a teacher must perform to do her job satisfactorily, but instead used liberal arts curricular documents to construct the entirety of the LAST-2.

[19] The Commission's report gives no indication that the Commission's determination stemmed in any way from a job analysis, or from the study of the job tasks New York teachers regularly perform.

In other words, NES started with the unproved assumption that specific facets of liberal arts and science knowledge were critically important to the role of teaching, and then attempted to determine how to test for that specific knowledge. This is an inherently flawed approach because at no point did NES ascertain, through an open ended investigation into the job tasks a successful teacher performs, whether its conception of the liberal arts and sciences was important to even some New York public school teachers, let alone to all of them. *See* Guidelines § 1607.5 ("Evidence of the validity of a test or other selection procedure by a content validity study should consist of data showing that the content of the selection procedure is representative of important aspects of performance on the job for which the candidates are to be evaluated.").

Dr. Buckendahl's Arguments to the Contrary Are Unpersuasive

Dr. Buckendahl, the BOE's expert, argues that NES's failure to identify job tasks does not necessarily render its job analysis unacceptable. He contends that because NES surveyed several hundred teachers about the importance of the KSAs that NES identified, and those teachers affirmed their importance, NES sufficiently demonstrated that those KSAs are necessary to the job of teaching. (Buckendahl Response 6–10). The Court finds these contentions unpersuasive, particularly given the high degree of validation required for tests, such as the LAST-2, that measure highly general abilities.[20]

[20] Additionally, there appears to be a critical difference between the methods by which Dr. Outtz and Dr. Buckendahl analyzed the LAST-2. Dr. Outtz states that his analysis relied most heavily on the Guidelines, *see* (Outtz Rebuttal 2–3), (March 20, 2015 Hearing Tr. at 41), while Dr. Buckendahl states that another authority, the *Standards for Educational and Psychological Testing*, "provides the most relevant guidance" for assessing employment examinations. (Buckendahl Response 2); Am. Educ. Research Ass'n, *Standards for Educational and Psychological Testing* (1999). Dr. Outtz claims that it is this difference that "underlies the difference in [their] opinions [as to the validity of the LAST-2] in this case." (Outtz Rebuttal 2). As the Supreme Court has made clear, and later Second Circuit decisions confirm, the Guidelines are "entitled to great deference" because they represent "the administrative interpretation of [Title VII] by the enforcing agency." *Griggs*, 401 U.S. at 433–34; *see also Gulino II*, 460 F.3d at 383; *Guardians*, 630 F.2d at 91. The same cannot be said for the *Standards*, which were formulated by the American Educational Research Association, and not executive branch officials. Accordingly, Dr. Buckendahl's substantial reliance on the *Standards* further undermines his conclusions.

The problem with NES's approach, and with Dr. Buckendahl's endorsement of it, is that it assumed, without investigation or proof, that specific KSAs are important to a teacher's effectiveness at her job—namely, an understanding of some pre-determined subset of the liberal arts and sciences—and then asked teachers to rank only those KSAs in importance. The fact that survey respondents stated that certain surveyed KSAs were important to teaching says nothing about the relative importance of the surveyed KSAs compared to any KSA not included in NES's survey. *See Guardians*, 630 F.2d at 95 ("[A] job analysis requires determination of the *relative importance* of the identified work behaviors." (emphasis added)). NES cannot determine the KSAs most important to teaching by surveying only those KSAs NES already believed were important. It must determine which KSAs to survey based on an investigation of the job tasks performed by successful teachers. Only KSAs which NES has directly linked to those identified job tasks should be included in a survey attempting to determine "relative importance." *See Vulcan Soc'y*, 637 F. Supp. 2d at 111 (finding that defendant's failure to provide evidence linking KSAs to job tasks "undermines the court's confidence that 'the pertinent abilities have been selected for measurement.'" (quoting *Guardians*, 630 F.2d at 96)).

As an example, and as a way of making these issues somewhat more concrete, assume that the KSA of reading comprehension has an importance value of 9, the KSA of logical reasoning has an importance value of 4, and the KSA of leadership has an importance value of 20. Assume that NES's survey would have queried the value of both reading comprehension and logical reasoning, but not of leadership. Ranked relative to each other, reading comprehension would be very important, while logical reasoning might be somewhat important. But in this example, neither is nearly as important as leadership. In this way, NES's survey would have greatly exaggerated the importance of both reading comprehension and logical reasoning.

This defect would not have existed had NES used an appropriate method to identify the job tasks of New York teachers in the first place. If leadership were important to the job of teaching, the identified job tasks would have made that clear, and the survey NES sent to educators would have included the KSA of leadership alongside the KSA of, for example, reading comprehension. This process would have provided NES with a much more accurate understanding of what KSAs are most important to the job of teaching, both overall and relative to one another.

NES's survey thus failed to ascertain what KSAs are most important to the job of teaching. Although this sort of survey might be an appropriate way of confirming information gathered through a proper job task investigation, or as a way of determining the relative importance of already-ascertained job tasks, it is not an appropriate way of initially identifying KSAs.

Additionally, Dr. Buckendahl's reliance on NES's surveying of educators is further undermined by the deficiency in the survey's sample. NES should have determined the size and construction of its sample by taking into account the makeup and number of all of the subgroups (e.g. kindergarten teachers, special education teachers, African-American teachers, New York City teachers) NES needed to survey in order to achieve a representative sample.[21] Doing so would have ensured that each of those subgroups was sufficiently represented in the responses

[21] According to experts in the field of organizational psychology, accounting for the views of different subgroups is a critical part of determining an appropriate sample size:

> [S]ample size should be determined by the number of factors that will be studied to determine their effect on the way work is perceived and performed. For instance, to investigate task difficulty for male and female employees by educational level and by location would require comparisons of task difficulty indices calculated separately for males and females at different locations and varying education levels. As the number of variables to be compared increases, so will the need for larger samples to enable those comparisons.

David M. Van De Voort, et al., *Work Analysis Questionnaires and App Interviews*, *in* The Handbook of Work Analysis 58 (Mark A. Wilson, et al. eds., 2012).

NES received. The small size of NES's overall sample, and the vanishingly small number of responses from African-American and Latino teachers—critical subgroups for ensuring the LAST-2 was unbiased and non-discriminatory—indicates that NES failed to do so.

As the Court discussed in Part IV, *supra*, very few of the survey respondents were African-American or Latino.[22] The SED argues that NES's survey was nonetheless demographically appropriate because the number of African-American and Latino teachers who responded to the survey is commensurate with the percentage of African-American and Latino educators in New York state, (SED Response 6): according to the SED, 10% of the respondents were African-American or Latino, which is not far from those groups' representation in the New York State teacher population in 2003–2004 school year (which was 4.8% African-American, and 7% Latino). (*Id.*) (citing *Schools and Staffing Survey*, Nat. Center for Education Statistics (last visited April 22, 2015), http://nces.ed.gov/surveys/sass/tables/state_2004_18.asp).

Although the percentages match approximately, the raw number of minority respondents—twenty-four African-American respondents and ten Latino respondents—was too small to permit NES to determine whether the answers received from minority teachers differed from those of majority teachers in any statistically meaningful way. (Outtz Report 33); *see also Exxon Corp. v. XOIL Energy Res.*, Inc., 552 F. Supp. 1008, 1021 (S.D.N.Y. 1981) (Broderick, J.) (taking issue with a survey because "[o]nly 97 respondents were interviewed in each of the two segments of the survey" and "all of the interviewees came from the vicinity of New York City," concluding that "the survey was not conducted on a properly selected and representative sample

[22] The SED admits in its response to Dr. Outtz's report that the database it used to determine whom to survey did not capture race or ethnicity information. (SED Response 6). Therefore, the SED and NES had no idea whether it was sampling a demographically appropriate population when it sent out its surveys. In the future, it will be necessary for the SED to either capture race and ethnicity information in the database it intends to use to survey educators so that it can ensure that all of its surveys are demographically appropriate, or it will need to supplement its initial mailing with additional surveys to ensure a proper sample.

of the population."); *cf. Vista Food Exch., Inc. v. Vistar Corp.*, No. 03-CV-5203, 2005 WL 2371958, at *7 (E.D.N.Y. Sept. 27, 2005) (rejecting a consumer survey after finding the sample of 75 respondents to be too small); *Bonechi v. Irving Weisdorf & Co., Ltd.*, No. 95-CV-4008, 1995 WL 731633, at *8 (S.D.N.Y. Dec. 8, 1995) (Schwartz, J.) (finding sample of 69 participants too small to adequately represent universe of potential consumers of New York City souvenir books). It was incumbent on NES to use a larger sample, and to consider whether it should oversample certain demographic groups that are small in number, to ensure that each of those groups is meaningfully represented in the survey. *See* Irwin L. Goldstein, et al., *An Exploration of the Job Analysis–Content Validity Process*, *in* Personnel Selection in Organizations 27 (Neal Schmitt & Walter C. Borman eds., 1993) ("[I]f there are groups of individuals within the organization that are small in number (perhaps ethnic minorities or women) . . . it usually makes sense to overrepresent these groups to ensure that the sample size is large enough to be representative of their views of the job and to provide an opportunity to collect sufficient data to determine whether there are differences in the way the job is viewed by members of these groups."); Irwin Goldstein & Sheldon Zedeck, *Content Validation*, *in* Fair Employment Strategies in Human Resource Management 32 (Richard S. Barrett ed., 1996) (noting that "it is important to represent the views of different groups of individuals, such as members of minority groups or women"); *cf.* Nathan D. Woods, *Assessing the Validity of Statistical Samples in Medicare Audits Key Items for Auditors and Providers to Consider*, 13 J. Health Care Compliance 71, 72 (2011) ("A common misperception in sampling is that the necessary size of a sample can be determined simply through considering the size of the population from which the sample is drawn. This is not the case. A far more important consideration is the degree of variation present in the data.").

The Procedure NES Should Have Followed

Because this is the second time during this case that NES has failed to complete properly a job analysis with respect to an employment exam, it may be useful to describe how a lawful job analysis should proceed.

NES should begin by first identifying the necessary job tasks for a New York public school teacher. Necessary job tasks could be identified through some combination of (1) teacher interviews, (2) observations of teachers across the state performing their day-to-day duties, *see* (Outtz Report 23–25) (discussing methods used for gathering job task information); *cf. Guardians*, 630 F.2d at 95 (approving of a portion of the defendant's job analysis, where "work behaviors involved in being a police officer were identified by extensive interviewing, and subjected to serious review"), and (3) the survey responses of educators who have been given open-ended surveys requiring them to describe the job tasks they perform and to rank the importance of those tasks, *see* (Outtz Report 23–25); *cf. Vulcan Soc'y*, 637 F. Supp. 2d at 111 (describing the use of "job questionnaires" to develop a list of job tasks). Simply consulting educational curricular documents is not a sufficient way of identifying job tasks or KSAs. Job tasks must be ascertained from the source—in this case, from public school teachers.

Using the data culled from such an investigation, NES could then analyze these job tasks, and from that analysis determine what KSAs a teacher must possess to adequately perform the tasks identified. *See* Guidelines § 1607.14(C)(4) ("For any selection procedure measuring a knowledge, skill, or ability the user should show that (a) the selection procedure measures and is a representative sample of that knowledge, skill, or ability; and (b) that knowledge, skill, or ability is used in and is a necessary prerequisite to performance of critical or important work behavior(s)."). NES should document precisely how those KSAs are necessary to the

performance of the identified job tasks. *See* Guidelines § 1607.15(A)(3). It is those KSAs that should provide the foundation for the development of the test framework.

The importance of identifying these job tasks is amplified here because every teacher in New York must be licensed, whether she teaches kindergarten, or advanced chemistry. *See* (Feb. 19, 2015 Hearing Tr. 18–19). NES therefore needs to determine exactly what job tasks are performed, and accordingly, what KSAs are required, to teach kindergarten through twelfth grade proficiently. This is likely a daunting task given how different the daily experience of a kindergarten teacher is from that of an advanced chemistry teacher.

Last, NES needs to make sure that the relevant test (here, the LAST-2) tests for abilities not already tested for by related exams. Here, applicants must also pass the ATS-W and the appropriate CST before they can become licensed.

ii. NES's Flawed Job Analysis Renders the Remainder of NES's Validation Procedure Deficient

A job analysis serves as the foundation for every other aspect of the validation process *Guardians* requires. NES's failure to perform a proper job analysis infected every other part of its validation process, rendering each similarly deficient.

Reasonable Competence. Testmakers are generally viewed as having used reasonable competence if the exam was created by professional test preparers, and if a sample study was performed that "ensure[d] that the questions were comprehensible and unambiguous." *M.O.C.H.A. Soc'y II*, 689 F.3d at 280. Here, NES, a professional test preparer, *see Gulino III*, 907 F. Supp. 2d at 519, conducted a sample study, *see* (Clayton Decl. ¶¶ 46–48). This showing is insufficient, however, when a portion of the test development process—in this case, the job analysis—is so wholly deficient. Such a pervasive error inherently negates what might otherwise

be a finding of reasonable competence. The LAST-2 thus fails to conform to the second *Guardians* factor.

Content Relatedness. Assessing the content relatedness of an exam "is intertwined with the job analysis." *Gulino III*, 907 F. Supp. 2d at 520. Content relatedness is demonstrated by showing that the "'abilities tested for . . . adequately relate[] to most of the identified tasks.'" *Vulcan Soc'y*, 637 F. Supp. 2d at 116 (quoting *Guardians*, 630 F.3d at 98). Because the law requires a job analysis to begin with the identification of job tasks, NES's failure to identify job tasks makes it impossible to assess the content-relatedness of the LAST-2.

Representativeness. For the same reasons, NES has also failed to demonstrate that the content of the exam is "a representative sample of the content of the job." *Guardians*, 630 F.2d at 98. The representativeness requirement has two components: "[t]he first is that the content of the test must be representative of the content of the job; the second is that the procedure, or methodology, of the test must be similar to the procedures required by the job itself." *Id.* Because NES never identified the tasks that make up the job, it is impossible to determine whether the content of the LAST-2 is representative of that job, or whether the test's procedures are similar to those of the job.

Scoring. Nor is it possible for the Court to determine whether the LAST-2's scoring system "usefully selects from among the applicants those who can better perform the job." *Guardians*, 630 F.2d at 95. Because NES did not define initially what the job of teaching entails, it is not possible to determine whether the scoring system used by the LAST-2 selects those applicants who can better perform that job.

The LAST-2 thus fails to meet any of the five criteria set forth in *Guardians* to assess whether an exam has "sufficient content validity to be used notwithstanding its disparate racial

impact." *Id.* Therefore, the Court finds that the LAST-2 was not properly validated and is not job related.

VI. CONCLUSION

For the reasons set forth above, the Court finds that the BOE violated Title VII by requiring Plaintiffs to pass the LAST-2 in order to receive a permanent teaching license. The parties shall submit a joint status letter to the Court by June 29, 2015, identifying what steps need to be taken in accordance with this Opinion.

SO ORDERED.

Dated: New York, New York
June 5, 2015

/s/
KIMBA M. WOOD
United States District Judge