ACT WORKING PAPER 2017-2

Understanding *Gulino*: A Legal Perspective

Michelle Croft, JD, PhD

April, 2017

ACT Working Paper Series

ACT working papers document preliminary research. The papers are intended to promote discussion and feedback before formal publication. The research does not necessarily reflect the views of ACT.

Michelle Croft is a principal research associate in the Office of Public Affairs.

The author would like to thank Dan Vitale for his helpful comments on earlier versions of this paper.

Understanding *Gulino*: A Legal Perspective

Gulino v. New York State Department of Education is a class action lawsuit that involves a teacher certification examination in New York. The case spanned over two decades and three iterations of the exam. The case is unique as the plaintiffs were already teaching in classrooms when they were required to sit for a new certification exam. Despite the unique nature of the case, it is a useful case study to examine what factors the court uses in evaluating employment assessments where there is a disparate impact (i.e., one racial/ethnic group consistently scores lower than another group).

This paper will first discuss the legal criteria used by courts in evaluating challenges to employment assessments, then describe the *Gulino* case and its resulting opinions.

Disparate Impact and Employment Assessments

There are a variety of ways an examinee may challenge an employment assessment through the courts when assessment results show that one racial/ethnic group performs differently than another on the assessment.[i]

Fourteenth Amendment of the U.S. Constitution

The Fourteenth Amendment of the U.S. Constitution requires that "No state shall make or enforce any law which shall abridge the privileges or immunities of citizens of the United States; nor shall any state deprive any person of life, liberty, or property, without due process of law; nor deny any person within its jurisdiction the equal protection of the laws." Actions under the Fourteenth Amendment related to employment assessments are typically related to equal protection (i.e., classifications or grouping) or due process.

A Fourteenth Amendment challenge is limited in reach. First, it requires a government or state action, so private employers are not included. Second, the action must affect a liberty or property interest. A teaching license is considered a property right (see *State Board of Education v. Drury*)[ii] and, therefore,

certain procedures must be in place (Vance, 1988). Related to due process, there must be notice and other procedures (such as retesting) to ensure that a license is not arbitrarily taken from a teacher. For equal protection, if the challenge involves racial/ethnic differences in score outcomes, then "strict scrutiny" applies. This means that the government must show that it has a substantial interest (e.g., ensuring that students are properly educated) and that the requirement is narrowly tailored (i.e., measuring the appropriate skills).[iii]

Title VII of the Civil Rights Act of 1964 (Pub. L. 88-352)

Title VII of the Civil Rights Act of 1964 was created to expand civil rights protections to employees. Title VII prohibits employment discrimination based on race, color, religion, sex, and national origin.[iv] Title VII is more extensive than the Fourteenth Amendment because it extends protections to employees beyond governmental actions. Specifically, it applies to employers who are engaged in an industry affecting commerce who have fifteen or more employees for each working day in each of twenty or more calendar weeks in the current preceding year. The statute allows for administrative remedies that are unavailable through the Fourteenth Amendment and requires a different type of legal analysis, which has two necessarily sequential components, discussed below.

1. Disparate Impact

Under Title VII, if an employment practice causes a disparate (or adverse) impact on the basis of race, color, religion, sex, or national origin, then the employer must demonstrate that the practice is job related for the particular position and consistent with business necessity (42 U.S.C. § 2000e-2(k)(1)(A)). Disparate impact is typically established using the Four-Fifths Rule where the pass rate (also called selection rate) is calculated for each racial/ethnic group, and impact ratios are calculated by comparing the pass rate for each group to the highest pass rate. There is evidence of disparate impact if the pass rate for any group is at most four-fifths or 80%of the pass rate for the highest group (EEOC, 1979).

Table 1. Four-Fifths Rule

Race/Ethnicity	# Examinees	# Passed	Pass Rate	4/5ths
White	100	75	75%	n/a
African-American	60	30	50%	50/75 = 67%
Latino	55	35	64%	64/75 = 85%

Table 1 provides two demonstrations of the calculation. For each racial/ethnic group, we calculate the pass rate. In this case, White examinees had the highest pass rate (80%). The next step is to compare the lower pass rates to the highest pass rate and determine whether each result yields an impact ratio that is at or below four-fifths. Thus the impact ratio for African-American examinees, 67%, would be evidence of disparate impact because it is lower than 80%, while the impact ratio for Latino examinees, 85%, would not be evidence of disparate impact because it is higher than 80%.

Because the four-fifths rule is a rule of thumb is based solely on outcomes, it is not consistent with the psychometric definition of fairness. The psychometric definition of fairness would include information about the test and item development process, differential item functioning, examinees' previous performance, or examinees' education history (e.g., course-taking information). The psychometric fairness information is relevant in the second portion of a Title VII analysis—job relatedness—but it is not considered as part of the disparate impact analysis.

2. Job Relatedness and the *Guardians* test

Even if there is evidence of disparate impact, the employer can continue to use an assessment if the employer can establish that the assessment is job related (42 U.S.C. § 2000e-2(k)(1)(A)). An assessment is job related if it has been properly validated by professionally acceptable methods (*Albemarle Paper Co. v. Moody*, 422 U.S. 405 (1975) at 431, quoting 29 C.F.R. 1607.4(C)) or if there is a manifest relationship to legitimate employment goals (*Griggs v. Duke Power Co.*, 401 U.S. 424, 432

(1970); *Watson v. Fort Worth Bank & Trust*, 487 U.S. 977 (1988)). If the employer is able to establish the validity of the assessment, the burden then shifts back to the plaintiffs to show that there is another equally efficient and trustworthy test or selection device that would serve the same legitimate purpose without the racial/ethnic effect (*Watson*, 487 U.S. at 998).

Individual jurisdictions may have their own specific tests to determine job relatedness that are consistent with U.S. Supreme Court precedent. The United States Court of Appeals Second Circuit, where *Gulino* was tried, uses the *Albemarle Paper* "proper validation" test. In describing the test, Second Circuit courts have acknowledged that validation is "not primarily a legal subject" (*Guardians*, 630 F.2 at 89; *Gulino V*, 907 F.Supp.2d 294 (S.D.N.Y. 2012)). Instead, a balance between two bodies of evidence is required: the expertise of test validation professionals, such as conformance with the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) (*Gulino IV*, 460 F.3d at 383 and *Gulino V*), and the Equal Employment Opportunity Commission's *Uniform Guidelines on Employment Selection Procedures* (29 C.F.R. 1607.1–1607.18). The *Uniform Guidelines* are not binding on the court, but they are entitled to deference (*Griggs v. Duke Power Co.*, 401 U.S. 424).

The Second Circuit developed a five-part test in *Guardians Ass'n v. Civil Service Commission of New York*, 630 F.2d 79 (2d Cir. 1980) to determine if an assessment has been properly validated for Title VII purposes. Part 1 is a job analysis. The test developer must: (1) identify the tasks involved in performing the job; (2) determine the relative importance of the tasks; and (3) identify the knowledge, skills, and abilities necessary to complete the tasks. The developer also must include a thorough survey of the relative importance of the skills and the degree of competency required for each skill.

Part 2 is reasonable competence in construction. Generally, an assessment developed by a professional developer will meet this requirement. Evidence that the reasonable competence requirement has *not* been met can include the lack of a pilot study to ensure that test items are comprehensive and unambiguous, or a pilot that is not representative of the tested population.

Part 3 is that the assessment content must be directly related to the content of the job. The *Gulino V* court highlighted that Part 3 is so related to the job analysis that if there are flaws in the job analysis, the employer must present convincing evidence that the assessment is job related.

Part 4 is similarly related to the job analysis and requires that the assessment content be representative of the content of the job (29 C.F.R. 1017.14(C)(4)). The employer must provide evidence of important knowledge, skills, and abilities required for the job and how they are tested on the exam. In doing so, the developer must ensure that the assessment primarily measures the important, instead of minor, aspects of the job.

Part 5 is related to scoring and requires that the scoring system usefully selects applicants who can better perform the job. The employer must present evidence that the cut score is reasonable and consistent with normal expectations of minimal qualifications or proficiency.

In examining the *Guardians* test, it is clear that that if the job analysis is not properly conducted, the assessment will not be "properly validated" under Title IV, as it would fail Parts 1, 3, and 4 of the test. Parts 2 and 5 are more consistent with the psychometric validity requirements found in the *Standards for Educational and Psychological Testing*.

Gulino Background

From 1986 to 1991, New York City awarded conditional licenses if candidates, depending on their status, completed either of two exams, "open" or "closed" (*Gulino III*, 2003 U.S. Dist. LEXIS 27325 (S.D.N.Y. 2003)). The open exam—a set of short-answer items, a test of written English, and an oral interview—was available to all teacher candidates who had or were in the process of attaining a college degree. The closed exam, which consisted solely of the oral interview, could be taken only by people who had worked as substitute teachers in the New York City public schools for two years under a "temporary per diem certificate."

In both cases, the licenses were conditional on the examinee completing additional New York City Board of Examiners ("the Board") requirements (such as completing additional coursework) within five years (*Gulino I*, 201 FRD 326 (S.D.N.Y. 2001)).[v]

In 1991, the state legislature eliminated the Board in favor of expanding a requirement to pass the National Teacher Core Battery (NTE)—a liberal arts assessment that measured knowledge of English, American history, math, science, art, and music (*Gulino I*)—that had been in effect since 1984 for all teachers in the state except those in New York City and Buffalo (*Gulino III* at 9–10). After the elimination of the Board, all teachers within New York were required to meet the state's credentialing requirements. Even so, a person could still teach in New York State without having passed the requirements for state certification if they were granted a temporary license, which could be renewed for up to three years (*Gulino III*). Because of the New York City teacher shortage, temporary licenses were repeatedly renewed.

In 1993, the State Department of Education (SED) replaced the NTE with the Liberal Arts and Sciences Test (LAST); thousands of teachers who had been credentialed by New York City between 1986 and 1991 failed the tests (i.e., the NTE and/or the LAST). The SED allowed those who failed to retake the assessments to attain qualifying scores rather than lose their jobs (*Gulino I*).

Gulino was brought in 1996. The plaintiffs were African-American and Latino teachers in the New York City public school system who had either lost their teaching licenses or were prevented from obtaining a full license because they did not pass either the NTE or the LAST (*Gulino II*). The plaintiffs alleged that the use of the certification exams to determine who would receive a permanent teaching certificate resulted in a disparate impact on African-American and Latino teachers in the New York City public schools, and was therefore in violation of Title VII of the Civil Rights Act of 1964 (§ 701 et seq., as amended, 42 U.S.C. § 2000e et seq.) (*Gulino II*, 236 F. Supp. 2d 314).

Initial statewide statistics for the NTE indicated that pass rates for White, African-American, and Latino teachers were approximately 84%, 40%, and 44%, respectively. For the LAST, pass rates were slightly higher: 93%, 50%, and 56%, respectively (*Gulino I*). For New York City teachers, pass rates were lower for all teachers than the state averages but were consistent in that the rates among White teachers were higher than those for the other groups.

Gulino generated more than 30 separate court opinions. Many of them deal with issues outside of test validity (e.g., certification of the plaintiff class, damages) and thus will not be dealt with here.

2003 Trial (*Gulino III*)

The eight-week trial in 2003 examined the validity evidence for both the LAST and the NTE. After hearing testimony, the court found that the defendants had sufficiently demonstrated only that the NTE was properly validated.

LAST Test Development (*Gulino III* and *V*)

National Evaluation Systems (NES), developer of the LAST, drafted a framework of five areas to be tested: (1) Scientific and Mathematical Processes; (2) Historical and Social Scientific Awareness; (3) Artistic Expression and the Humanities; (4) Communication Skills; and (5) Written Analysis and Expression. Within each area, four or five subtopics would be tested. A Bias Review Committee and Content Advisory Committee reviewed the draft framework and subtopics. NES conducted a survey of New York state public school teachers and college faculty members asking them to rate the importance of each subtopic to public school teaching, then developed an item bank of 350 items reflecting the survey results. The Bias Review and Content Advisory Committees reviewed 80 of the 350 items. NES then administered a pilot exam of 36 items to students at New York education colleges. The final version of the exam consisted of an essay, 64 operational multiple-choice items that counted toward the student's score, and 16 multiple-choice field test items that did not count. The operational multiple-choice portion represented 80% of the total score, while the essay represented the remaining 20%.

Standard setting was conducted by the Bias Review and Content Advisory Committees, who reviewed an early version of the LAST to estimate the average essay score and the difficulty level of each multiple-choice question (the percentage of examinees that would answer the question correctly). The committees also recommended passing scores (also known as cut scores) for the exam, with the Content Advisory Committee recommending a higher score (48) than did the Bias Review Committee (38). The state Commissioner of Education used the lower score. In 1997, a new committee recommended increasing the cut score to 44; the Commissioner raised it to 43.

The trial court opinion described a pervasive lack of documentation with respect to the LAST. The developer, NES, had not kept documentation related to the content domains to determine whether they were developed prior to or after the task analysis was performed. Comment forms from the panelists who reviewed the objectives were missing and notes or records from the Equity Advisory Board were absent. Similarly, NES could not identify teachers, content specialists, and editors involved in item writing. In addition to documentation problems, the court also highlighted issues in setting the cut score. NES used a Modified Angoff method (in which experts examine item content and predict how many items a minimally qualified examinee will answer correctly) but did not use impact statistics after reviewing the proposed cut score when it was initially set or when it was subsequently raised. Further, the panel only considered a selection of items when recommending the cut score.

Despite finding insufficient validation evidence for the LAST, the court held, following *Watson v. Fort Worth Bank & Trust* (487 U.S. 977), that the LAST was "manifestly" (i.e., on its face) related to legitimate employment goals. The court found that the LAST was job related primarily because of the essay portion of the exam. The results of a survey of education professionals indicated that the ability to write an essay was an important skill for teachers. Given the weight of the essay, the court found that the majority of plaintiffs would have passed the LAST except for the essay portion. Since the essay portion was job related, the court found that the assessment did not violate Title VII.

LAST Appeal (*Gulino IV*)

Both the plaintiffs and the defendants appealed the trial court's ruling. Even though the defendants were not liable, they appealed the decision that they could be subject to Title VII liability as they did not want the precedent established.[vi] The plaintiffs appealed the decision that the LAST was job related and faulted the court for failing to address their argument that the Board misused the NTE and the LAST by requiring experienced teachers to take the exams.

The appeals court found that the trial court had made a legal error related to the validity analysis by applying the requirement in *Watson* that only a manifest relationship need prove job relatedness. The appeals court cited *Albemarle Paper* (422 U.S. 405, 431) which states that "discriminatory tests are impermissible unless shown, by professionally acceptable methods, to be predictive of or significantly correlated with important elements of work behavior which comprise or are relevant to the job or jobs for which candidates are being evaluated." The appeals court held that the district court should have used the five-part test in *Guardians* described above rather than *Watson* to determine the validity of the LAST.

The appeals court also found that the trial court had made two factual errors related to the validity analysis. First, the appeals court ruled that a lack of documentation does not necessarily preclude an assessment from meeting the validation standard so long as there are first-hand accounts and expert testimony (p. 388). Second, the appeals court found that there was insufficient evidence to support the trial court's conclusion that the majority of plaintiffs would have passed the LAST except for the essay writing portion (p. 386).

LAST Remand (*Gulino V*)

Because the appeals court found fault in the trial judge's decision, the case was returned to the trial court for remand. As before, the remand court found that the NTE was properly validated. However, in applying the five-part *Guardians* test rather than *Watson*, in 2012 the court found that the LAST was not properly validated.

Under *Guardians*, the main deficiency was Part 1 the job analysis. The remand court found that NES did not conduct a suitable job analysis to ensure that the exam adequately assessed the knowledge, skills, and abilities needed for daily job tasks (*United States v. City of New York*, 637 F.Supp. 2d 77, 111 (E.D.N.Y. 2009)).The remand court also found that there were flaws in the method NES used to develop and review the subtopics that rendered the job analysis unsuitable: NES did not create a list of tasks teachers perform nor determine whether the subtopics identify the knowledge needed to perform the tasks, nor was there evidence that NES had identified important tasks or assessed the relative importance of tasks. Finally, the remand court ruled that there was insufficient evidence regarding the materials NES used to draft subtopics. As mentioned above in the trial discussion, NES was unable to produce evidence regarding interviews with faculty, teachers, and content experts and could not establish how the interviews were incorporated into the subtopics.

Part 2 of the *Guardians* test—reasonable competence—was also found not to have been met. The remand court highlighted gaps in test construction and flaws in the pilot study, such as relying on a subset of items, and only including college students when the exam was also intended for working teachers. The court also noted that NES did not retain documentation of the construction process.

Because the job analysis was inadequate, the remand court found that the content of the LAST was not directly related to teaching and thus did not satisfy Part 3 of the *Guardians* test. Although the LAST was related to liberal arts and sciences broadly, that fact, the court ruled, was insufficient for establishing job relatedness. Instead, there needed to be evidence of what minimum knowledge *about* liberal arts and sciences would ensure competency for all teachers.

Similarly, the court found that the representative content requirement (Part 4 of the *Guardians* test) was not met, because the lack of a sufficient job analysis meant that the Board could not establish that the LAST only measures important aspects of the job.

Last, with respect to Part 5 of the *Guardians* test, the remand court found flaws in the standard setting such that there was not sufficient evidence that scoring requirements usually selected individuals who would be better teachers. Only a subset of items—80 of 350—were reviewed during the standard setting, and there was not an independent review of the remaining items. The court also noted that it was not clear whether the Bias Review and Content Advisory Committees used proper criteria to set the cut score. The committees were asked to imagine the "minimally competent teacher" without any guidance regarding the definition or types of teachers that might meet the standard. There were no notes from the standard setting, and committee members were not asked to explain their rationales. One member of the Bias Review Committee who served on the score review committee testified that it was not the purpose of the standard setting to distinguish between minimally competent and incompetent teachers—leading the court to conclude that the panel was not properly trained. Finally, the court faulted the Commissioner of Education for raising the cut score despite statistics showing that, given the lack of a demonstrable relationship between LAST scores and teacher performance, raising the score would disproportionately affect racial/ethnic minority examinees.

LAST-2, 113 F.Supp. 3d 663 (S.D.N.Y. 2015)

Test Development

The LAST was updated between 2000 and 2004 and was renamed the LAST-2. NES added Technical Processes and Research Skills to the assessment framework, which otherwise remained similar to the earlier framework for the LAST (henceforth "LAST-1"). NES then developed objectives and made each objective into "focus statements" that provided details about the content. The new framework, objectives, and focus statements were based both on the LAST-1 framework and on documents describing common liberal arts and science course requirements in New York state colleges and universities.

The LAST-2 framework was again reviewed by the Bias Review and Content Advisory Committees. NES then conducted two separate surveys of New York educators, asking respondents to

rate the importance of each objective to public school teaching. The survey was sent to 500 teachers and 320 responded. Twenty-four of the respondents were African-American and 10 were Latino. The second survey was sent to 181 college faculty members and 45 responded. Of the college respondents, none were African-American and three were Latino.

The SED approved the LAST-2 Framework, and NES began item development. Some of the items were repurposed from existing LAST-1 item bank. New items were reviewed by the Bias Review and Content Advisory Committees. The Content Advisory Committee also reviewed the repurposed items. For pilot testing, some of the new or repurposed items were embedded on the LAST-1 and others were administered to volunteer examinees.

The cut score was set by a panel of New York educators who were asked to imagine a hypothetical individual who was just at the level of knowledge and skills required to perform as a teacher in New York, and then to determine the number of items that individual would answer correctly.

Disparate Impact

As with the LAST-1, pass rate was higher for White examinees than for African-American and Latino examinees (54% and 75% of the White rate, respectively). Because there was disparate impact and because the LAST-2 was related to the previously contested LAST-1, the court retained authority over it and evaluated the LAST-2 to determine if it was properly validated.

Job Relatedness

As with the LAST-1, the court found that the LAST-2 was not properly validated and that this was primarily due to the job analysis. The court stated that instead of starting with the premise that teachers should demonstrate an understanding of the liberal arts, the starting point should have been looking at job tasks performed by New York teachers. The court found that NES did not identify any job tasks, and therefore could not determine the relative importance of each task as required by Part 1 of the *Guardians* test.

The court also found that the educator survey was insufficient, for a number of reasons. First, the surveys assumed that the knowledge, skills, and abilities (KSAs) were important to a teacher's job and restricted respondents to ranking the importance of only those KSAs, without permitting them to identify others. Further, the court linked this insufficiency to those of the job analysis, stating (as per *United States v. City of N.Y.*, 637 F.Supp. 2d at 111) that job tasks should have been used rather than KSAs. Finally, given the low response rates for African-American and Latino educators, the survey sample was insufficient.

Because the LAST-2 represented the second time that the *Guardians* test was not met, the court explicitly laid out the procedures that NES should have followed. As mentioned above, the court stated that the first step is to identify the necessary job tasks to be a New York public school teacher. This could be done through teacher interviews, teacher observations of day-to-day duties, and a survey of responses that includes open-ended items requiring them to describe the job tasks they perform and rank the importance of those tasks. The second step is to use the data collected to analyze the job tasks to determine what KSAs a teacher must have to perform those tasks. In doing so, the developer must take into account what tasks are required for *all* teachers, because the test is administered in all content areas. Finally, the developer should ensure that the content is not being tested on related exams.

ALST, 122 F.Supp. 3d 115 (S.D.N.Y. 2015)

The LAST-2 was replaced with the Academic Literary Skills Test (ALST) in May 2014, which was also a measure of literacy skills. The change was in response to the federal Race to the Top competitive grant program, which rewarded state and district reform in K-12 education. In 2010, New York adopted new pedagogical and curricular standards that "redefined" the teacher's role. The ALST was intended to measure "A teacher candidate's literacy skills . . . reflecting the minimum knowledge, skills, and abilities an educator needs to be competent in the classroom and positively contribute to student learning." Teacher candidates were also required to pass the Educating All Students test (EAS), which measured

skills and competencies related to diverse student populations such as English learners and students with disabilities, and the edTPA, which focused on pedagogy.

Test Development

As part of Race to the Top, the state Board of Regents directed the SED to develop the New York Teaching Standards that outlined the requirements for certification and the expectations of teachers throughout their careers. The Teaching Standards were revised using a working group of educators and parents and were subsequently released for public comment. Each Teaching Standard is defined at three levels of specificity: standard, element, and performance indicator.

SED developed the ALST Framework by identifying the KSAs that the SED thought an incoming New York state public school teacher must successfully perform. The test developer, Pearson, based the Framework on the Teaching Standards and the New York State P–12 Common Core Learning Standards for English Language Arts and Literacy ("Common Core"). Pearson identified two KSAs: "Reading" and "Writing to Sources" and the Framework included "Performance Indicators" with additional detail. The Framework was reviewed and updated by the SED and the Bias Review and Content Advisory Committees.

Pearson then conducted a content validation survey of 500 New York public school teachers asking about the importance of the KSAs for performing their job. They also asked how well each of the Performance Indicators represented important examples of the KSAs and how well the KSAs represented important aspects of the knowledge and skills needed to teach. Two hundred twenty-three teachers completed the survey; 3.5% of respondents were African-American and 9.1% were Latino. Pearson analyzed the results separately by racial/ethnic group and found that African-American and Latino teachers had rated the KSAs and Performance Indicators similarly. The survey was also sent to 112 teacher preparation faculty members, and 63 surveys were returned. None of the faculty respondents were African-American or Latino.

The Human Resources Research Organization (HumRRO) was subcontracted to conduct the job analysis, which occurred after Pearson developed the ALST Framework. The job task analysis was completed by reviewing the Teaching Standards, Common Core, other standards documents developed by state or national organizations, academic articles discussing teaching practices, and the O*NET online job task database. The initial list included 101 tasks divided into seven categories.

A Job Analysis Task Force composed of New York public school educators and education preparation program faculty reviewed the list of job tasks, yielding a final list of 105 tasks across the seven categories. HumRRO then administered a survey of New York educators asking them to rate the importance of the job tasks and indicate whether any job tasks were omitted. Over 7,000 educators were sent the survey and 1,655 completed it. HumRRO analyzed the data by looking at different teaching assignments and developed a formula to identify critical tasks that were frequently performed. They identified 34 tasks that were rated as critical to teaching.

The Job Analysis Task Force was also asked to assess the importance of the Performance Indicators for "Reading" and "Writing to Sources." Finally, a focus group conducted a "linkage exercise" to link the Performance Indicators to job tasks.

For item development, Pearson created an assessment specification document to provide guidelines on reading passage selection and item writing. The items were reviewed by the Bias Review and Content Advisory Committees. There were two pilots, again via embedded field test items and administration to a volunteer group.

Standard setting was performed by a panel of 18 subject matter experts, including two African-Americans. As before, the panel used a Modified Angoff method to recommend a cut score, to the Commissioner of Education and the SED.

Disparate Impact

There was a dispute over whether there was a disparate impact for the ALST. Because the court determined that the exam was job related (see below), it did not make a finding on whether or not there was a disparate impact.

Job Relatedness

Unlike in the previous cases, the court found that the ALST met all of the requirements of the *Guardians* test. Regarding Part 1, it was job related, in that the SED properly relied on the various standards documents and that the Teaching Standards and Common Core can be considered written job descriptions. Although the documents did not represent what teachers were doing at the time, they did represent what teachers were going to be expected to do in the future as part of the Race to the Top reform. The documents went into detail on how and what teachers should teach, and the Common Core specifically included the literacy standards that were required for all teachers.

Regarding Part 2, the court found that there was reliable competence in constructing the ALST, particularly because all of the items were field tested, not just a subset. Regarding Part 3, the court ruled that the content of the ALST was related to teaching. There was testimony that literacy skills are critical in teaching. The ALST Framework included these through the "Reading" and "Writing to Sources" KSAs and their accompanying Performance Indicators, which were used to create the assessment specifications. Finally, the material was reviewed by the Content Advisory Committee.

Regarding Part 4, the court ruled that the content of the ALST was representative of a teacher's job, in that the survey of teachers indicated that more than half of a teacher's daily job involves literacy skills. Finally, regarding Part 5, the court ruled that Pearson had used an accepted method for standard setting.

Nevertheless, the court did note a couple of flaws in the ALST development process. The survey samples and focus groups did not adequately represent the population of New York public school

teachers. Also the job tasks should have been linked to the Performance Indicators rather than to the KSAs as HumRRO had done. But the court found that this was not invalidating because the Performance Indicators were clearly linked to Common Core State Standards and this link was sufficient to ensure job relatedness.

Despite meeting Title VII requirements, New York Board of Regents decided to discontinue the ALST assessment in March 2017 due to differences in passing rates for minority examinees (Taylor, 2017). One of the reasons for doing so was that deans of education schools argued that the exam was "exacerbating a shortage of teachers of color" within the New York public schools.

Conclusion

The *Gulino* case provides an overview of Title VII requirements that can be useful for test developers and states using these types of assessments. As the cases demonstrate, disparate impact is different from the psychometric conception of fairness. Title VII uses evidence of disparate impact as a prima facie case of discrimination, which is a fairly blunt tool. After disparate impact has been established, then the courts look more at the typical test development and validity evidence called for in the *Standards for Educational and Psychological Testing*, which are applicable to all testing programs, and at the job analysis. For assessments potentially challenged under Title VII, the quality of the job analysis is critical. Without a proper job analysis, the assessment is unlikely to meet the Title VII requirements.

References

Albemarle Paper Co. v. Moody, 422 U.S. 405 (1975)

Debra P. v. Turlington, 564 F.Supp, 177 (M.D. Fla. 1983), *affirmed*, 654 F.2d 1079 (5th Cir. 1981).

Griggs v. Duke Power Co., 401 U.S. 424, 432 (1970)

Guardians Ass'n v. Civil Service Commission of New York, 630 F.2d 79 (2d Cir. 1980)

Gulino v. Bd. of Educ. of the City Sch. Dist. of New York, 201 F.R.D. 326 (S.D.N.Y. 2001) ("*Gulino I*")

Gulino v. Bd. of Educ. of the City Sch. Dist. of New York, 236 F. Supp. 2d 314 (S.D.N.Y. 2002) ("*Gulino II*"), modified, 2002 U.S. Dist. LEXIS 24965, 2002 WL 31887733 (S.D.N.Y. 2002).

Gulino v. Bd. of Educ. of the City Sch. Dist. of New York, 96 Civ. 8414, 2003 U.S. Dist. LEXIS 27325 (S.D.N.Y. 2003) ("*Gulino III*")

Gulino v. Bd. of Educ. of the City Sch. Dist. of New York, 460 F.3d 361 (U.S. Court of Appeals 2d Cir. 2006) ("*Gulino IV*")

Gulino v. Bd. of Educ. of the City Sch. Dist. of New York, 907 F.Supp.2d 492 (S.D.N.Y. 2012) ("*Gulino V*")

Gulino v. Bd. of Educ. of the City Sch. Dist. of New York, 113 F.Supp. 3d 663 (S.D.N.Y. 2015) ("*LAST-2*")

Gulino v. Bd. of Educ. of the City Sch. Dist. of New York, 122 F.Supp. 3d 115 (S.D.N.Y. 2015) ("*ALST*")

Phillips, S.E. (2010). Assessment law in education, available at https://www.dokshop.com/MainHome.asp

State Board of Education v. Drury, 263 Ga. 429, 437 S.E.2d 290 (1993)

Taylor, K. (2017, March 13). Regents drop teacher literacy test seen as discriminatory. *The New York Times*, March 13, 2017, https://www.nytimes.com/2017/03/13/nyregion/ny-regents-teacher-exams-alst.html

Title VII of the Civil Rights Act of 1964 (§ 701 et seq., as amended, 42 U.S.C. § 2000e et seq.)

The U.S. Equal Employment Opportunity Commission (EEOC). (1979). Adoption of questions and answers to clarify and provide common interpretation of the Uniform Guidelines on Employee Selection Procedures, *Federal Register* 44(43), March 2, 1979.

Vance, Chris E. (1998). Teacher competency testing: "Decertification" and the federal constitution and Title VII, 37 Emory Law Journal 1077.

Watson v. Fort Worth Bank & Trust, 487 U.S. 977 (1988)

[i] For a detailed discussion related to legal challenges to assessments in the area of education more broadly, see Phillips (2010).
[ii] In the K–12 context, a high school diploma is considered a property right (*Debra P. v. Turlington*).
[iii] If the challenge does not involve race, then a lower level of scrutiny—rational basis—is used. The rational basis standards only requires that the government have a legitimate interest and that there is a reasonable relationship.
[iv] Title VI of the Civil Rights Act prohibits discrimination in race, color, and national origin in programs, such as education institutions, that receive federal assistance. Title VI has been used to challenge assessment programs that result in a disparate impact (e.g., *GI Forum et al. v. Texas Education Agency et al.*).
[v] If teachers who passed the open exam did not meet the additional requirements within five years, they could still have their City license restored if they later met the requirements; whereas those who had passed the closed exam could not have their license restored (*Gulino III*).
[vi] The appellate court ultimately dismissed claims against SED finding that it was not an employer, but held that Title VII applied to the Board because it was both a licensor and the plaintiffs' employer.