IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF DELAWARE

UNITED STATES OF AMERICA, Plaintiff, v. THE STATE OF DELAWARE, THE DELAWARE DEPARTMENT OF PUBLIC SAFETY, and THE DELAWARE DIVISION OF STATE POLICE, Defendants.	)	Civil Action No. 01-020-KAJ

POST-TRIAL FINDINGS OF FACT AND CONCLUSIONS OF LAW

Patricia C. Hannigan, Esquire, United States Attorney's Office, Wilmington, Delaware, plaintiff, United States of America.

Michael W Tupman, Esquire, Department of Justice, Wilmington, Delaware, counsel for defendants.

David H. Williams, Esquire, Morris, James, Hitchens & Williams LLP, Wilmington, Delaware, Wayne S. Flick, Esquire and Robert J. Malionek, Esquire, Latham & Watkins, Los Angeles, California, counsel for defendants, the State of Delaware, the Delaware Department of Public Safety, and the Delaware Division of State Police.

March 22, 2004
Wilmington, Delaware

TABLE OF CONTENTS

I. INTRODUCTION 1

II. FINDINGS OF FACT 1

A. The Hiring Process at the DSP 2

1. The use of the Alert test 2

2. Other aspects of the hiring process 5

3. The four-phase probationary period for newly hired Troopers 7

B. The Importance of Reading and Writing Skills for the Trooper Job 8

C. Assessing the Validity of the Alert in Measuring Reading and Writing Skills 10

1. Reliability and content validity 11

2. Criterion-related validity 16

D. The Parties' Efforts to Determine a Cutoff Score on the Alert That Adheres to the *Lanning* Standard 25

1. Utility and expectancy tables 25

2. Dr. Wollack's two-step analysis 30

3. False positives and false negatives 36

a. False positives 36

b. False negatives 38

4. Regression analysis 42

a. Reverse regression 43

b. Forward regression 51

5. Character of the Trooper job 52

III. CONCLUSIONS OF LAW 54

JORDAN, District Judge

I. INTRODUCTION

The United States brought this employment discrimination action against the State of Delaware, the State's Department of Public Safety, and that department's Division of State Police (collectively the "State" or "DSP" or the "Defendants"), pursuant to Section 707 of Title VII of the Civil Rights Act of 1964, as amended, 42 U.S.C. §§ 2000e-6, et seq. (*See* Docket Item ["D.I."] 1.) In an earlier Opinion, I held that the United States had established a *prima facie* case that the Defendants' use of a multiple-choice reading comprehension and writing test known as the "Alert" to screen applicants seeking employment as DSP Troopers had a disparate impact on African American applicants because those applicants passed the Alert at a statistically significantly lower rate than Caucasian test takers. (D.I. 261.) A bench trial was held from August 13 to August 20, 2003, to afford the Defendants an opportunity to demonstrate that, despite the disparate impact of the Alert test, their use of that test from 1992 to 1998 was lawful because it was "job related for the position in question and consistent with business necessity." *See* 42 U.S.C. § 2000e-2(k)(1)(A)(i). As required by the Third Circuit's opinions in *Lanning v.Southeastern Pennsylvania Transportation Authority* ("SEPTA"), 181 F.3d 478 (3d Cir. 1999) (*Lanning I*), and *Lanning v. SEPTA*, 308 F.3d 286 (3d Cir. 2002) (*Lanning II*), the standard by which the Defendants' use of the Alert is to be judged is, whether the discriminatory cutoff scores applied by the Defendants in screening applicants with the Alert measured "the minimum qualifications necessary for

successful performance of the job" of DSP Trooper. *See Lanning I*, 181 F.3d at 489. I have concluded that the Defendants have failed to meet their burden of proof and that, while the Alert is a valid and reliable test for law enforcement employment screening, the Defendants set the cutoff score at an impermissibly high level. I have further concluded that the range within which the cutoff score could reasonably have been set is 66 to 70%.

The following post-trial findings of fact and conclusions of law are issued pursuant to Federal Rule of Civil Procedure 52(a).[1]

II. FINDINGS OF FACT

A. The Hiring Process at the DSP

1. The use of the Alert test

1. From 1981 through October 1998, Defendants used the Alert as part of their entry-level Trooper selection process. (D.I. 263 at p. 3, ¶ 1.)[2] The United States challenges Defendants' use of the Alert as part of the selection process for recruit classes designated as Class 61 through Class 69. The time period at issue covers November 21, 1991 through October 1998. (*Id.* at p. 3, ¶¶ 1, 4.) After Class 69, the

[1]Throughout these Findings and Conclusions, I have adopted without attribution language suggested by one side or the other in this dispute. In all such instances, the Finding or Conclusion in question has become my own, based upon my review of the evidence and the law.

[2]D.I. 263 is the Final Pretrial Order and contains a recitation of uncontested facts.

Defendants replaced the Alert with another test.[3] (*Id*. at 263 at p. 3, ¶ 3; Tr. Vol. 3, 725:21-726:2.)

2. The Alert is a 160-item multiple choice test consisting of 60 items designed to measure reading comprehension and 100 items designed to measure four aspects of writing skills, namely, spelling, clarity, grammar, and detail. (D.I. 263 at p. 3, ¶ 2.) There are seven alternate forms of the test. (*Id*.)

3. The Alert's reading comprehension items require the test taker to read a passage and answer multiple choice questions based on the passage. The writing skills items require the test taker to choose the correct spelling of a word from among three choices, to choose the most clearly written of three statements, to choose the more grammatically correct of two sentences, and to choose which of three statements provides the most appropriate level of detail. (Ex. 61; Ex. 224 at p. 2; Tr. Vol. 1, 156:23-157:20.)

4. According to Dr. Stephen Wollack, a principal of Wollack & Associates, Inc. ("Wollack & Associates")[4] and the creator of the Alert test, the reading and writing skills assessed by the Alert are two aspects of a single ability called "prose literacy." (*See* Tr. Vol. 1, 157:21-158:6; Ex. 224 at p. 43.) As used by Dr. Wollack, the terms "reading and writing skills," "prose literacy," and "verbal ability" all refer to the same

[3]The United States has taken no position with regard to the lawfulness of Defendants' use of the replacement test. (D.I. 263 at p. 3, ¶ 3.)

[4]Wollack & Associates develops and validates tests for screening candidates for law enforcement employment. (Tr. Vol. 1, 8:23-9:19.)

thing – the reading comprehension and specific writing skills the Alert is meant to measure. (Tr. Vol. 1, 158:24-159:9.)

5. During the period at issue, Trooper applicants were required to pass the Alert and meet all other qualifications for employment.[5] (D.I. 263 at p. 3, ¶ 4.) The hiring process is highly selective. Out of more than 4500 applications received during that period, the DSP hired only 269 Troopers.[6] (Tr. Vol. 3, 726:21-727:4.)

6. For the recruit classes in question, the DSP used Alert cutoffs that range from 115 to 123, or 71.875% to 76.875%, varying by difficulty of test form. (D.I. 263 at pp. 3-5, ¶¶ 4, 5-14 & 27.) When Alert scores were standardized, the sample-size weighted cutoff score used during the period at issue was approximately 75% of items correct. (Tr. Vol. 6, 1617:15-1619:17; Ex. 205 at p. 36.)[7]

7. It is undisputed that the Alert assesses reading and writing skills that are relevant to the job responsibilities of a DSP Trooper. (D.I. 263 at p. 6, ¶ 35; Tr. Vol. 5, 1321:5-12; D.I. 301 at p. 2, ¶ 1; D.I. 304 at p. 1.)[8] It is also undisputed that the reading

---

[5] *See* ¶ 9, *infra*, for details on the other qualifications for employment as a DSP Trooper.

[6] On April 5, 2002, the parties filed a set of stipulated facts which, among other things, identify each applicant who took the Alert during the period at issue, his or her race and Alert score, and whether he or she passed or failed the Alert. (D.I. 263 at p. 4, ¶ 15; D.I. 120.)

[7] Although slightly different raw score cutoffs were used on different Alert forms, 75% is the score used for ease of reference by the parties (*see* D.I. 301 at p. 2, n. 2; Ex. 205 at p. 36) and in this decision.

[8] D.I. 301 is the Defendants' proposed findings of fact and conclusions of law; D.I. 302 is the United States' proposed findings of fact and conclusions of law. D.I. 303 and D.I. 304 are, respectively, the Defendants' objections to D.I. 302 and the United States' objections to D.I. 301.

and writing demands on entry level law enforcement officers such as DSP Troopers are much the same throughout the United States. (D.I. 263 at p. 5, ¶26.) The parties also agree that the reading and writing skills measured by the Alert are only part of a broad range of skills required for effective service as a DSP Trooper. (*See* Ex. 208 at p. 8; Tr. Vol. 5, 1321:10-21.)

8. Those who failed the Alert were ineligible to continue in the hiring process for that recruit class, but could take the Alert again the following year. (D.I. 263 at p. 4, ¶ 17.)

2. Other aspects of the hiring process

9. The hiring process employed additional steps that attempted to assess other skills and qualities important for service as a DSP Trooper. If an applicant passed the Alert, he or she was then required to move through those additional steps, including the following:

i. During the period at issue, the selection process required that applicants for the DSP Trooper job have a high school diploma or GED and at least 60 semester or 90 quarter credit hours from an accredited college or university, equivalent to an associate's degree. (D.I. 263 at p. 4, ¶ 16.)

ii. The selection process included use of the Police Attitudinal Factors examination developed by Wollack & Associates and used to assess an applicant's attitudes in five areas, namely, race relations, use of force, use of authority, flexibility, and maturity. (D.I. 263 at p. 4, ¶ 18.)

iii. The selection process also included use of the Personal History Questionnaire, which questioned an applicant about his or her background, or the

Lifestyle Examination, later renamed the Disclosure Statement, which consisted of a series of questions pertaining to the minimum qualifications for the position, criminal activity, work experience, and various attitudinal factors. The answers to the Personal History Questionnaire and the Lifestyle Examination/Disclosure Statement were later confirmed by interview or background investigation. (D.I. 263 at p. 4, ¶ 18.)

iv. The selection process included the use of an oral interview, with questioning by a board of five DSP officers. The board members independently rated each applicant in five categories: attitude, appearance, communication skills, fairness, and decision making. (D.I. 263 at p. 5, ¶ 19.)

v. In some years, the selection process included the use of a writing sample. (D.I. 263 at p. 5, ¶ 20.) When such a sample was used, the DSP reviewed and considered it during the final selection stage of hiring. (Tr. Vol. 3, 742:9-743:23.) Writing samples were never used to eliminate a candidate. (Tr. Vol. 3, 743:17-19.)

vi. The selection process included the use of a physical fitness test to assess an applicant's aerobic capacity, muscular strength and endurance, and flexibility. (D.I. 263 at p. 5, ¶ 21.)

vii. The selection process included the use of a polygraph examination to assess an applicant's truthfulness in responding to questions regarding the applicant's use of aliases or incorrect names, education record, marital and personal relationships, permanency intentions, employment records, debts, accident and traffic violation record, arrests or participation in undetected crimes, illegal use of drugs, subversiveness, gambling, and alcohol consumption. (D.I. 263 at p. 5, ¶ 22.)

viii. The selection process included the use of a background investigation designed to reveal whether an applicant was suitable for employment in light of his or her demonstrated character traits and past behavior. (*See* D.I. 263 at p. 5, ¶ 23.)

10. Applicants who satisfactorily completed the Defendants' pre-offer selection process were considered for conditional offers of employment. Those applicants were then required to complete a medical history and submit to a medical examination including a physical and laboratory testing, an eye examination, a physical fitness assessment, and a psychological evaluation. (D.I. 263 at p. 5, ¶¶ 24-25.)

3. <u>The four-phase probationary period for newly-hired Troopers</u>

11. Applicants who were hired embarked upon a two-year probationary period and participated in a preparatory training program divided into four phases. (Tr. Vol. 3, 705:2-13; Ex. 68 at pp. DELMS 3854-3855). In Phase I, Troopers attended the DSP Training Academy for about twenty-two weeks, during which each Trooper's performance was evaluated through written tests and daily observations. (Tr. Vol. 3, 680:23-681:13; 705:14-706:10.) At the conclusion of Phase I, DSP Troopers were required to take and pass the Delaware Council on Police Training ("COPT") certification test. (D.I. 263 at p. 5, ¶ 28; Tr. Vol.3, 682:20-683:9; 706:11-707:3; D.I. 302 at p. 3, ¶ 9.)

12. Troopers who completed Phase I and passed the COPT test became eligible for Phase II. Phase II was a twelve-week field training and evaluation program (the Field Training Officer, or "FTO", Program). (D.I. 263 at p 6, ¶ 29.) During the FTO program, Troopers were rated daily on twenty-seven dimensions of job performance.

(Tr. Vol. 3, 708:21-23.) By the end of Phase II, a Trooper had to have achieved a minimally acceptable rating in each of the twenty-seven areas to be eligible for Phase III. (Tr. Vol. 3, 707:4-709:17; 711:19-715:18; D.I. 302 at p. 4, ¶ 10.)

13. Phase III was a six-month period during which each Trooper was monitored and evaluated monthly by supervisory personnel. (Tr. Vol. 3, 709:18-710:16.)

14. Phase IV of the preparatory training program was the Trooper's second year of employment, during which a Trooper remained in a probationary status and was evaluated quarterly. (Tr. Vol. 3, 710:17-711:18; D.I. 302 at p. 4, ¶ 11.)

B. The Importance of Reading and Writing Skills for the Trooper Job

15. It is crucial for Troopers to read and write well in order to fulfill their role as protectors of public safety. Investigating and reporting unlawful activity is at the core of their responsibilities. In our complex society, those responsibilities demand literacy at a level that addresses both the need to conduct investigations according to evolving legal standards and the need to accurately communicate the results of an investigation.[9]

---

[9]The United States repeatedly emphasized in its post-trial briefing the many other aspects of a Trooper's job besides literacy. While that point has relevance (*see infra* at ¶ 43 and n. 29), the emphasis given to it was largely misplaced. It is true that the skills measured by the Alert do not represent the full range of skills needed to perform the job of Trooper. (Tr. Vol. 5, 1321:10-1322:24.) The Alert measures a part of the reading and writing domain, which is a subpart of the verbal domain, which is a subpart of the cognitive domain, which is a subpart of the DSP Trooper job universe. (Tr. Vol. 5, 1334:23-1337:19.) But while I have no doubt that physical fitness, psychological stability, cultural sensitivity, personal integrity, and other qualities of mind and body are of great importance in a Trooper's work, the only issue before me is whether the Alert was properly used in screening applicants. No one asserts, nor could it be credibly asserted, that literacy is unimportant for success as a Trooper. A functionally illiterate officer will not be successful, no matter how fine and accomplished an individual he or she may be in other respects. I therefore do not see the question as one of balancing

16. In the course of conducting investigations, Troopers must read and apply a great deal of written material, including legal manuals, the Motor Vehicle and Criminal Codes, law updates issued by the Attorney General's office, court decisions, and protection from abuse orders, and they must do so while on the road (*e.g.*, on their mobile computers while responding to a complaint), at home, or at the office.[10] (Tr. Vol. 3, 782:23-786:6.) Troopers also read background information in case files in order to prosecute misdemeanors in the Justice of the Peace Courts. (Tr. Vol. 3, 796:14-797:8.) Much of the material Troopers must read and apply frequently, such as the Standard Operating Procedures specific to each Troop and the 380-page DSP Administrative Manual, which outlines policies for responding to various situations, is not taught at the Academy and is updated often. (Tr. Vol. 3, 751:11-754:16; Ex. 253.) The DSP gives Troopers on-the-job training throughout their careers to help them stay abreast of changes in Delaware's criminal and motor vehicle codes, as well as developments in the law of evidence and constitutional law. (*See* Tr. Vol. 3, 688:11-694:2.) In short, Troopers must be able to read, understand, and apply on a daily basis information from a variety of sources, much of which is abstract and intellectually challenging. (*See* Tr. Vol. 3, 688:11-691:21.)

17. Troopers also dedicate a significant portion of each day to writing reports about their investigations. (*See* Tr. Vol. 3, 694:9-20.) They write reports while on

---

the qualities called for in the job. I am called upon to reach a conclusion about the minimum level of literacy for the job of DSP Trooper, so literacy and the use of the Alert to measure it are the focus of this decision.

[10]Troopers typically have their offices at a location called a "Troop." The term "Troop" also refers to a division or group of Troopers within the DSP.

patrol, while at the Troop, and, sometimes, while at home. (Tr. Vol. 3, 786:7-9; 791:13-22.) The timeliness and accuracy of their reports is critical, since the reports serve an essential function in the administration of justice. If a matter becomes contested, they are a record that will be referred to again and again, and they will rightly be subjected to searching inquiry by private litigants, by prosecutors and defense attorneys, by probation officers and judges, and by the press and public. That which goes unreported, or which is reported in an inconsistent or incoherent way, may be treated as fiction, and the resulting disservice to the facts may also result in a serious disservice to the interests of justice. (*See* Tr. Vol. 3, 791:23-792:11.) As one Trooper said during the trial, "[i]f it's not in the report, it's not taken as credible." (Tr. Vol. 3, 791:21-22.)

C. Assessing the Validity of the Alert in Measuring Reading and Writing Skills

18. There is basic agreement between the parties that literacy is an essential aspect of a Trooper's job. The parties also agree that the Alert assesses relevant literacy skills. (*See supra* at ¶ 7.) There is, however, vigorous debate over the degree of validity of the assessment yielded by the Alert and over the cut-off score appropriate to establish that Trooper candidates have the minimum level of literacy necessary for successful performance as a Trooper.

19. One of the ways to demonstrate that a test such as the Alert is an appropriate screening device is through a statistical validation study.[11] In the context of

[11]The use of statistical validation studies for such a purpose is described in Chapter 5 of the treatise *The Statistics of Discrimination*, by R. Paetzold and S. Willborn (West 2002). The Third Circuit has noted the role that validation studies can play in determining whether an employment practice is job-related and consistent with business necessity. *See Lanning I*, 181 F.3d at 486 (quoting discussion in *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 431 (1975), regarding study of relationship between a test and

employment selection, a validation study essentially involves the establishment of a relationship between a selection procedure and a job or job performance. Two sets of professional standards, the American Psychological Association's *Standards for Educational and Psychological Testing* (1999), and the Society of Industrial and Organizational Psychology's *Principles for the Validation and Use of Personnel Selection Procedures* (1987), recognize that a selection procedure may be validated by content or by criterion-related methods. (Tr. Vol. 1, 24:10-30:11.) Content validity explains the extent to which the content of a test matches a particular job domain – that is, a set of abilities required for the job. (Tr. Vol. 1, 25:12-19.) Criterion-related validity explains the extent to which a selection instrument predicts a criterion, such as job performance. (Tr. Vol. 1, 29:8-30:8.) Neither method of validation is, in the abstract, superior to the other. (Tr. Vol. 1, 31:15-17.) *See also* 29 C.F.R. § 1607.5(A).

1. Reliability and Content Validity

20. At the trial in this case, the Defendants first presented evidence through Dr. Wollack, who is an expert in industrial and organizational psychology. (Tr. Vol. 1, 12:4-11.) Dr. Wollack has spent nearly thirty years developing employment tests for law enforcement officers and conducting validation studies of such tests. (Tr. Vol. 1, 9:8-19.) As noted earlier, *see supra* at ¶ 4, he is the creator of the Alert.

21. Dr. Wollack was retained by the State to conduct a validation study of the Alert in Delaware and to evaluate the DSP's Alert cutoff scores, which, despite an obvious degree of self-interest on his part, was a reasonable decision, given that Dr.

"important elements of work behavior," using "professionally acceptable methods").

Wollack has already conducted several validation studies of the Alert in other locales. (Tr. Vol. 1, 40:18-41:20; Exs. 224, 225, 226.) Dr. Wollack also provided testimony regarding the reliability of the Alert as a selection measure.

22. One method for determining the reliability of an employment test like the Alert is to measure its content validity. (Tr. Vol. 1, 25:7-11.) Content validity is the extent to which the content of a test "matches," or corresponds to, the set of related abilities that are required to perform a certain job. (Tr. Vol. 1, 25:14-19.)

23. Dr. Wollack testified that content validity can be established by either direct or indirect methods. (Tr. Vol. 1, 27:7-21.) Content validity is established directly when a test representatively samples job tasks or behaviors. (Tr. Vol. 1, 26:24-27:4.) It is established indirectly when a test measures skills and abilities that are necessary to perform the job. (Tr. Vol. 1, 27:18-21.) The indirect method of establishing content validity requires two steps, first, proving that the test accurately measures what it purports to measure, and, second, showing that the skills measured by the test are necessary and important for performing the job. (Tr. Vol. 1, 27:22-28:7.)

24. In this case, Dr. Wollack did not rely on the direct method of showing content validity. (Tr. Vol. 1, 185:12-15; Tr. Vol. 5, 1355:20-1357:6.) Rather, Dr. Wollack sought to determine whether the Alert reliably measures reading and writing skills, and whether reading and writing skills are important and necessary for the Trooper job. (Tr. Vol. 1, 35:11-20.) Thus, using the indirect method, Dr. Wollack testified that the Alert is content valid because Troopers need to read and write and the Alert is a reading and writing test. (Tr. Vol. 1, 184:8-18.) In other words, the test measures skills necessary for the job of DSP Trooper. (*See* Tr. Vol. 1, 35:11-20.)

25. Since its development in 1976, the Alert has been the subject of several validation studies. (Ex. 224 at p. 4.) Dr. Wollack's expert report identifies 20 such studies conducted between 1982 and 2001, including 6 content validation studies and 14 predictive validation studies.[12] (*Id*.) The content validation studies included two statewide studies, one conducted in Texas in 1990 and one conducted in Washington in 1991. (Exs. 240, 241, 242; Tr. Vol. 1, 50:7-51:5.) A total of 82 police departments have participated in the content validation studies of the Alert. (Ex. 224 at p. 4.)[13]

26. Reliability of a test is a necessary condition for validity. *See* Paetzold & Willborn, *supra* at n. 11, at § 5.12. Relying on his previous studies of the Alert, Dr. Wollack concluded that the Alert reliably measures the reading and writing skills required to perform the entry-level law enforcement job.[14] (*See* Tr. Vol. 1, 39:2-40:21; Ex. 224 at pp. 9-12.) A retest reliability estimate[15] for the Alert was computed by the

---

[12]The predictive validation studies were undertaken to determine the degree to which the Alert is correlated with the performance of recruits in police training academies. (Ex. 224 at p. 4.)

[13]Dr. Wollack testified about the value of having multiple validation studies of the Alert. (Tr. Vol. 1, 42:20-43:21.) He emphasized that an ongoing research program over a period of years and across many jurisdictions permits the confirmation of research findings in different locations and the pooling of data, both of which assist in achieving reliable estimates. (*Id*.) As previously noted, the parties agree that the reading and writing demands of the entry-level law enforcement job are much the same from jurisdiction to jurisdiction. (D.I. 263 at p. 5, ¶ 26.)

[14]Reliability studies provide important information about the consistency of a person's scores on a series of measurements. (Ex. 224 at p. 9.) The reliability coefficient (reported as r_{xx}) is the statistical index by which the degree of test reliability is expressed. (*Id*.) The reliability coefficient may vary in magnitude from 0 to 1, with 0 representing no reliability and 1 representing theoretically perfect reliability. (*Id*.)

[15]Retest reliability assesses the similarity of scores for a group of people in two applications of the same test. (Ex. 224 at p. 9.)

Washington State Criminal Justice Training Commission, with a sample size of 633 job applicants who retested with the examination.[16] The resulting retest reliability was r_{xx}=.90. (Ex. 224 at pp. 9-10.) Internal consistency reliability estimates[17] from the Missouri State Highway Patrol, City of Janesville, Wisconsin, Hartford, Connecticut, Hawai'i County, Hawaii, and the Minnesota Department of Public Safety, were averaged to arrive at a resulting internal consistency reliability estimate of r_{xx}=.93, from a sample of 4,344 applicants. (*Id*.) These studies also demonstrate parallel forms reliability[18] among the Alert forms used by the State. (Ex. 224 at pp. 9-12.)

27. I am persuaded that the Alert is reliable in the technical, statistical sense.

28. In order to again assess the validity of the Alert, Dr. Wollack first assessed the Trooper job in Delaware. In doing so, he worked with Subject Matter Experts ("SMEs"),[19] including incumbent entry-level officers and supervisors. (Tr. Vol. 1, 60:8-

---

[16]In their proposed findings of fact, the Defendants relied upon the sample size being 633 job applicants. (D.I. 301 at p. 10, ¶ 28.) However, there is a discrepancy in the sample size in the exhibit upon which the Defendants rely. In a table reflecting the reliability estimates of the Alert, the sample size is indeed 633 (Ex. 224 at p. 10), however, in the narrative describing how the retest reliability estimate was computed, it states that the sample size was 644. (Ex. 224 at p. 9.) Regardless, the computation with either sample size appears to have resulted in a reliability coefficient equal to .90. (Ex. 224 at pp. 9-10.)

[17]Internal consistency reliability refers to the degree of test homogeneity – that is, consistency in the way in which the test takers respond to the test questions. (Ex. 224 at p. 9.)

[18]Parallel forms reliability estimates the similarity of scores on different forms or versions of a test. (Ex. 224 at p. 9.)

[19]SMEs are experienced job incumbents who assist psychologists in understanding the abilities that are required to perform a given job. (Tr. Vol. 1, 61:4-11.)

91:12; Ex. 224 at pp. 13-50.) A "Job Analysis Panel," consisting of a cross-section of DSP Officers from the rank of entry-level Trooper to Captain, compiled a list of a Trooper's job tasks and a list of the skills and abilities required to perform those tasks. (Tr. Vol. 1, 60:16-62:17; Ex. 224 at pp. 13-27.) Dr. Wollack also collected job analytic data from entry-level Troopers and supervisors through surveys. (Tr. Vol. 1, 84:13-17.) Supervisors reported that reading and writing are among the most important skills for assessment in an entry-level selection process. (Tr. Vol. 1, 85:15-88:18; Ex. 244 at pp. 46-48; Ex. 285.) Dr. Wollack concluded that DSP Troopers routinely depend upon written materials to perform essential tasks, that report preparation is an important and frequent part of the job, and that reading and writing pervade the job. (Tr. Vol. 1, 90:15-91:12; Ex. 224 at pp. I, 69.) This Delaware finding is consistent with Dr. Wollack's findings in studies in Missouri, Washington, Texas, and Colorado. (Tr. 85:15-88:18; Ex. 224 at pp. 46-48; Ex. 285.)

29. Dr. Wollack's study also included readability analyses that showed that the reading level of the Alert matches the reading level required for the DSP Trooper job. (Tr. Vol. 1, 79:8-84:7; Ex. 224 at pp. 57-59.) His finding in this regard was corroborated by results in previous studies. (Tr. Vol. 1, 82:6-84:7; Ex. 224 at pp. 57-59.)

30. Dr. Wollack's past studies of the Alert, as well as his study specific to the DSP, led him to conclude that the Alert is valid as a job-related measure of prerequisite reading and writing skills required for the Trooper job in Delaware. (*See* Tr. Vol. 1,

90:15-91:12.) His testimony and the evidence he relied upon were persuasive on this point, although the degree of validity was not meaningfully quantified.[20]

2. Criterion-related Validity

31. The State also presented evidence through Dr. P. Richard Jeanneret, an industrial organizational psychologist with more than 30 years of experience in developing and validating employee selection procedures. Dr. Jeanneret is the Managing Principal of Jeanneret & Associates, a Houston, Texas consulting firm that specializes in human resource management. (Tr. Vol. 2, 305:9-22.) He has conducted more than 200 validation studies, many of those in the law enforcement and public safety context. (Tr. Vol. 2, 311:23-313:16.) Dr. Jeanneret also has substantial expertise in designing methods for assessing job performance. (Tr. Vol. 2, 313:19-316:7.)

32. Dr. Jeanneret was retained by the State to conduct a criterion-related validity study of the Alert as it is used by the DSP, to examine the fairness of the Alert,[21] and to evaluate the DSP's Alert cutoff scores. (Tr. 326:16-327:24; Exs. 205 & 208.)

---

[20]Dr. Goldstein's testimony regarding the content validity of the Alert was not persuasive. He generally denigrated the use of multiple-choice tests in employee selection (*see* Tr. Vol. 5, 1373:1-1375:13), but it is clear that the professional standards in industrial psychology recognize the use of such tests for that purpose. (*See* Tr. Vol. 1, 154:13-155:12; Tr. Vol. 2, 480:19- 481:14.)

[21]"Test fairness," as distinguished from adverse impact, relates to whether a given test predicts job performance equally for members of different groups. When a test is determined to be fair, a common regression line describes the relationship between test scores and job performance for majority and minority group members. For example, if a white and African-American applicant correctly answer 86% of the items on a fair test, then the same level of job performance would be predicted for both. (Tr. Vol. 2, 338:8-340:3.)

33. Criterion-related validity involves a statistical analysis of the relationship between a predictor (in this case, the Alert) and a criterion (in this case, Trooper job performance). (Tr. Vol. 2, 327:4-16.) A criterion study determines whether a statistical relationship exists and, if so, the degree of confidence that can be placed in that relationship. (*Id*.) Criterion-related validity evidence provides a basis for drawing inferences from test scores, including inferences about predicted job performance. (*Id*.)

34. Dr. Jeanneret worked with a panel of six SMEs from the DSP to identify and define the various performance dimensions that make up the Trooper job. (Ex. 205 at p. 6.) The SME panel included three lieutenants, two sergeants and one captain. (*Id*.) The initial draft of the performance dimensions was based on job analytic information from the following sources: (1) data collected by Wollack & Associates, (2) data collected by an independent testing firm, SHL Landy Jacobs, that designed a new Trooper selection process for the DSP in 2000, (3) the DSP's existing performance appraisal process, (4) published literature concerning the police officer job, and (5) Jeanneret & Associates' own body of job analysis information. (Tr. Vol. 2, 342:12-343:24; Ex. 205 at p. 3.) The SME panel modified the initial draft, leading to the following list of 13 dimensions: oral communication, written communication, analyzing and problem solving, attention to detail, planning and organizing, adaptability and flexibility, judgment and decision-making, initiative and effort, integrity and professional commitment, interpersonal relations, stress tolerance, physical ability, and overall job knowledge. (Tr. Vol. 2, 345:10-346:4; Ex. 205 at pp. 7-8.)

35. Once the list of performance dimensions was finalized by the SME panel, Dr. Jeanneret and the SMEs created a Performance Dimension Rating Form (the

"PDRF"), which is simply a rating scale used as a performance evaluation tool. (Tr. Vol. 2, 346:5-9; Ex. 205, Appx. A.) For each performance dimension, a rating form was created that included five boxes. (Tr. Vol. 2, 346:22-347:6.) Three boxes were labeled "Outstanding," "Expected," and "Poor," and included examples of behaviors that the SME panel believed described performance at each level. (*Id*.) An unlabeled box was placed between the "Outstanding" and "Expected" boxes and another was placed between the "Expected" and "Poor" boxes. (Tr. Vol. 2, 346:5-347:21; Ex. 205, Appx. A at pp. 9-22.) These five boxes then served as rating categories, creating a 1-to-5 rating scale. (Tr. Vol. 2, 369:6-10.) Each PDRF rating form also included a 1- to-60 scale, with five groups of 12 numbered lines corresponding to each of the five boxes, such that lines 1 to 12 corresponded with box 1 ("Poor"); lines 13-24 corresponded with box 2; lines 25 to 36 corresponded with box 3 ("Expected"), and so on. (Tr. Vol. 2, 368:14-369:17; Ex. 205, Appx. A at pp. 9-22.)

36. A group of 62 supervisor/SMEs – all sergeants in the DSP – was assembled and provided with written instructions on completing the PDRF. (Tr. Vol. 2, 348:22-349:4; Ex. 205, Appx. A, pp. 1-3). The SMEs also received training by DSP Captain John Yeomans, whom Dr. Jeanneret had trained in the rating process. (Tr. Vol. 2, 350:21-351:15.) Each of the 62 DSP supervisor/SMEs rated each Trooper they supervised on each of the 13 dimensions. (Ex. 205 at p. 9 and Appx. B.) As a result, every DSP Trooper was rated. (Tr. Vol. 2, 351:16-352:6.) Each SME first assigned a Trooper to one of the five boxes ("Outstanding," "Expected," "Poor," or one of the in-between boxes), depending upon the Trooper's observed performance on each dimension. (Tr. Vol. 2, 373:5-16.) The SMEs then ranked each Trooper on the 1-to-60

scale corresponding to the broader boxes into which they had been placed. (Tr. Vol. 2, 737:17-24.) This process forced the SMEs to provide relative rankings for any two or more Troopers assigned to the same performance category. (Tr. Vol. 374:21-375:20.) As a consequence, the PDRF provided more refined performance information, as each Trooper whose performance was rated received a score on the 1-to-5 scale and a score on the 1-to-60 scale. (Tr. Vol. 2, 369:11-14.)

37. The SMEs were never asked to rank the Troopers for minimally acceptable performance. (Tr. Vol. 5, 1517:14-1518:10.) Dr. Jeanneret testified that, "[i]t's just never a terminology we've ever used." (Tr. Vol. 5, 1517:18-19.) Instead, using the terminology of "Outstanding, "Expected," and "Poor," experts for both the United States and the Defendants adopted the "Expected" rating as defining the level of performance that is the baseline of minimum qualification for Trooper success, as required by the *Lanning* standard. (*See* Tr. Vol. 5, 1514:21-1518:10.)

38. When the rating process was completed, Dr. Jeanneret examined the statistical relationships between the Alert scores Troopers received when they applied to the DSP and their PDRF performance ratings. Dr. Jeanneret first hypothesized that Alert scores would be significantly related to performance in oral communication, written communication, analyzing and problem solving, and attention to detail. (Tr. Vol. 2, 360:6-361:13.) He then correlated Troopers' Alert scores[22] with (1) their ratings on the

[22]Some Trooper candidates took the Alert multiple times. Approximately 10% of the Alert scores in Dr. Jeanneret's database were below the DSP's cutoffs because some Troopers failed the Alert at least once before obtaining a passing score and being hired. (Tr. Vol. 2, 357:2-14.) Dr. Jeanneret correlated those Troopers' first and last Alert scores with the various PDRF ratings. (Tr. Vol. 2, 357:22-358:3.) He found little difference in the statistical relationships, whether first or last Alert scores was used. (Tr.

1-to-5 scale for each of 13 performance dimensions; (2) their ratings on the 1-to-60 scale for each dimension; and (3) a composite of the 4 hypothesized dimensions (oral communications, written communication, analyzing and problem solving, attention to detail), which he labeled the "PDRF Composite." (Ex. 205 at p. 19.) Dr. Jeanneret found that Alert scores were statistically significantly related to the PDRF Composite on the 1-to-5 and 1-to-60 scales.[23] Those correlations are set forth below:

Vol. 1, 358:4-18; Ex. 205 at p. 19.)

[23]Last Alert scores were also statistically significantly related to performance in planning and organizing, adaptability and flexibility, and interpersonal relations. (Ex. 205 at p. 20, Table 8.)

Correlations Between Standardized First Alert Scores And PDRF Job Performance Ratings

	Alert Scores with 1-5 PDRF Scale	Alert Scores with 1-60 PDRF Scale	Alert Scores with 1-5 PDRF Scale corrected for Range Restriction	Alert Scores with 1-60 PDRF Scale corrected for Range Restriction	Alert Scores with 1-5 PDRF Scale corrected for Range Restriction and Criterion Unreliability	Alert Scores with 1-60 PDRF Scale corrected for Range Restriction and Criterion Unreliability
Oral Comm.	.21**	.21**	.27	.27	.33	.33
Written Comm.	.23**	.25***	.30	.32	.36	.39
Analyzing & Problem Solving	.20**	.19**	.26	.25	.31	.30
Attention to Detail	.16*	.15*	.21	.20	.25	.24
PDRF Composite	.24**	.23**	.31	.30	.37	.36

Notes: *p<.05, **p<.01, ***p<.001; these so-called "p-values" are measures used in judging statistical significance.[24] No test of statistical significance applies to the corrected validity coefficients in the above table.

(Ex. 205 at p.20, Table 8; Ex. 208 at p. 33.)

[24]The magnitude or size of a correlation is indicated by a numerical coefficient (such as .25 for the correlation between Alert scores and performance in written communication set forth in the table above). The confidence a researcher places in the relationship is expressed by the level of statistical significance. A benchmark level of statistical significance used by statisticians and industrial psychologists is .05, meaning that there exists a 1 in 20 probability that the observed relationship occurred by chance. (*See* Tr. Vol. 4, 1049:13-15.) More stringent levels of statistical significance are the .01 level (indicating a 1 in 100 probability that the observed relationship occurred by chance) and the .001 level (indicating a 1 in 1,000 probability that the relationship occurred by chance). (Tr. Vol. 4, 1264:4-1266:22.)

39. Dr. Jeanneret testified that, based on decades of research, these correlations indicate the relationship that one would expect between a test of cognitive abilities (such as the Alert) and performance in a law enforcement job. (Tr. Vol. 2, 381:1-387:24; 396:23-397:13; 401:8-404:10.) Dr. Harold W. Goldstein, the United States' expert on industrial psychology, agreed, stating that a well-known analysis involving hundreds of criterion-related validity studies showed that the correlation between tests of cognitive ability and law enforcement job performance ranged from .10 to .20.[25] (Tr. Vol. 5, 1412:5-1413:24.) Dr. Jeanneret observed correlations that fall within and slightly above that range.

40. However, as the Defendants concede, Dr. Jeanneret's reported correlations, noted in the table above, at most explain that performance on the Alert predicts between approximately 4% and 9% of the variance in the PDRF Composite ratings. (D.I. 301 at p. 19 n.17.) The smaller a correlation coefficient, the less power a test has to predict job performance. (*See* Tr. Vol. 4, 1040:2-1042:10; 1058:8-1059:15; Tr. Vol. 5, 1543:15-18.) The degree of prediction may be calculated by squaring the correlation coefficient. (Tr. Vol. 2, 386:14-24.) The resulting figure, called the "proportion of variance," represents the amount of variation in the predicted variable – in

---

[25]At trial, Dr. Goldstein testified that this range, from .10 to .20, was the "low to moderate" range of correlations. (Tr. Vol. 5, 1413:1-24; 11414:19-1415:3.) However, as discussed *infra* at n. 28, the statistical text cited at trial states that correlations of .1 are described as low and correlations of .3 are described as moderate. (*See* Tr. Vol. 2, 402:4-11; Tr. Vol. 3, 654:9-655:4; Tr. Vol. 4, 1042:11-1043:10 (citing a standard statistical text by Dr. Cohen).) Because none of the correlations developed by Dr. Jeanneret are .3 or greater, Dr. Goldstein's testimony is unpersuasive to the extent that it characterizes the range of correlations as moderate. (*See* Tr. Vol. 5, 1414:19-1415:3.)

this case, job performance – that is explained by the test score. Thus, for example, a correlation coefficient of .21 between the Alert and the PDRF job dimension of Oral Communication means that 4.4% of the variance in individuals' Oral Communication ratings can be explained by differences in their performance on the Alert.[26] (*See* Tr. Vol. 2, 386:4-24; Tr. Vol. 4, 1053:12-1059:15; Tr. Vol. 5, 1541:24-1542:24.)

41. Weak though the predictive capacity may be, however, if the strength of a statistical relationship is such that it reaches a benchmark level of statistical significance, then, as Dr. Bernard Siskin, the United States' expert statistician stated, one can conclude that the relationship between the two variables studied is "real." (Tr. Vol. 5, 1268:7-14.) Two of the correlations Dr. Jeanneret observed between Alert scores and performance on the dimensions that make up the PDRF Composite are statistically significant to the .05 level of significance, using a "one-tailed" test.[27]

---

[26]Thus, the calculations are as follows: $(.21)^2 = 0.0441$, and $(0.0441) \times (100) =$ 4.4% variance.

[27]In the argot of statistics, "one-tailed" and "two-tailed" tests of statistical significance refer to different ways of looking at the fundamental question of whether something observed is purely a matter of chance. (*See supra* at n.24.) The D.C. Circuit explained the difference between one- and two-tailed tests in *Palmer v. Shultz*, 815 F.2d 84, 92-96 (D.C. Cir. 1987). The "tails" refer to the ends of the well-known statistical bell curve that represents a random normal distribution. 815 F.2d at 93. In any random distribution, "the area under any segment of the bell curve measures the probability of that range of results occurring randomly." *Id*. In a one-tailed test, one looks at the random possibility at only one end of the curve; in a two-tailed test, one looks at both ends. This has practical implications because the assertion that an observed relationship between two variables is statistically significant is often made in employment cases by noting that the observed relationship reaches the .05 level of significance, i.e., that there is a 5% (or 1 in 20) chance that the relationship observed is purely random, or, put differently, that five percent of the area under the bell curve is in play. *Cf*. *id*. at 94-95 (describing .05 standard and the effect of a one-tailed or two-tailed test on meeting that standard). Of course, it makes a difference if that 5% is all at one end of the curve or is split between both ends of the curve, since the ends represent two

Furthermore, seven correlations are significant to the .01 level, and one is significant to the .001 level.

42. The evidence demonstrates that the relationship between Alert scores and performance in the relevant areas of the Trooper job is relatively weak but still provides an appropriate basis for decision-making by the State. In other words, the Alert has

---

different extremes in the relationship between the two variables. Dr. Siskin agreed with Dr. Jeanneret that ".05 is the normal standard chosen" (Tr. Vol. 3, 1049:15), but that does not answer the question of whether the correlations found by Dr. Jeanneret should be measured using a one-tailed or a two-tailed test. To say a one-tailed test is appropriate, one must assume, as Dr. Jeanneret did, that there will only be one type of relationship between the variables. (*See* Tr. Vol. 3, 1048:15-23.) Here, the rational assumption is that, if there is a problem, it is one of underselection of African-Americans, not overselection. Therefore, the probability of a chance deviation of African-American selection rates from selection rates for whites is properly measured by a one-tailed test, as Dr. Siskin seemed to concede. (*See* Tr. 1047:2-1050:6.) I therefore accept as persuasive the one-tailed test for statistical significance employed by Dr. Jeanneret. *See* Paetzhold & Willborn at § 4.14, p. 42 (supporting use of one-tailed test in discrimination cases because "[t]he question really being asked is whether the employer is behaving unfairly to ... [the plaintiff class] in its hiring process."); *but see Palmer*, 815 F.2d at 95 (stating that a two-tailed test should be applied because, in that gender discrimination case, "the hypothesis to be tested ... should generally be that the selection process treated men and women equally, not that the selection process treated women at least as well as or better than men.").

generally low criterion validity[28] but its predictive power is statistically significant. (Ex. 205 at pp. 19-20; Tr. Vol. 5, 1267:16-1269:12.)

D. The Parties' Efforts to Determine a Cutoff Score on the Alert That Adheres to the *Lanning* Standard

1. Utility and expectancy analyses

43. Having determined that the evidence establishes that the Alert has both content and criterion validity, although the degree of content validity is not quantified and the degree of criterion validity is relatively low, I turn next to the question of whether the cutoff score set by the Defendants fairly approximated the minimum literacy qualifications necessary for successful performance of the job of DSP Trooper. Dr. Jeanneret attempted to answer that question in part by conducting utility and expectancy analyses. A utility analysis is an "estimation of the institutional gains or losses anticipated from different employee selection strategies." (Ex. 205 at 35.) In this case, Dr. Jeanneret endeavored to show "changes in utility that result from increasing or decreasing cutoff scores on Alert ... ." (*Id*.) Unfortunately, the utility analysis here is of negligible value. While it purports to measure the marginal utility of a particular cut-off

[28]As discussed *supra* at n. 25, Whether the correlation between the Alert and performance should be characterized as "low" or "moderate" is a matter of earnest contention between the parties. (*See* D.I. 302 at p. 11, ¶¶ 35-40.) In a standard statistical text cited at trial, correlations of .1 are described as "low" and correlations of .3 described as "moderate". (*See* Tr. Vol. 2, 402:4-11; Tr. Vol. 3, 654:9-655:4; Tr. Vol. 4, 1042:11-1043:10.) The evidence shows that the correlations developed by Dr. Jeanneret are generally low and, indeed, in the only figures which can be shown to have statistical significance, namely the uncorrected correlations, none of them is .3 or higher. Nevertheless, given the statistical relationships that Dr. Jeanneret found between the Alert and the four dimensions making up the PDRF Composite, Dr. Siskin testified that, "I agree with him, there's evidence of validity." (Tr. Vol. 5, 1272:17-1273:6.)

score as a selection device, it does not give any meaningful answer to the question before me, namely what "discriminatory cutoff score measures the minimum qualifications necessary for successful performance of the job in question[.]" *Lanning I*, 181 F.3d at 489. The utility analysis seems only to support the unremarkable proposition that, the higher the score on the Alert, the more likely it is to screen out more candidates who might otherwise have difficulty performing as a Trooper, at least in the literacy aspects of the job.[29] But the "more is better" rationale in setting cutoff scores has been specifically rejected by the Third Circuit, *Lanning I*, 181 F.3d at 493, and I decline to follow the logic of the utility analysis to that conclusion. Even Dr. Jeanneret acknowledged that the utility analysis is "an index of how valuable the test was, but it would only be one piece of information. We might then want to look at selection rate. We might want to look at ... any number of things ... before we made a decision in terms of where to set the cutoff score." (Tr. Vol. 2, 410:12-19.)

[29]As Dr. Siskin aptly noted in his July 2002 rebuttal report (Ex. 8), filed in response to Dr. Jeanneret's April 2002 expert report, Dr. Jeanneret's utility analysis rests on two unproven and doubtful assumptions. First, the analysis assumes that the overall utility or value of a Trooper is directly related to his performance on the four abilities rolled into the PDRF Composite. Those literacy qualities, however, are focused on a single dimension in a multi-dimensional job. To use Dr. Siskin's analogy, there is no logical justification for assuming that a baseball player with superior fielding skills is necessarily a better overall player than one with weaker fielding skills, unless one also knows something about the player's ability in other critical dimensions of the job, such as hitting and running. (*See id*. at 7.) Second, Dr. Jeanneret assumed that increases in the PDRF Composite were associated with specific dollar amounts, so that a marginal value in dollars could be related to cut-off scores. The dollar valuations are based entirely on assumption, and, in any event, serve only to emphasize the "more is better" point, which flows from any direct, linear relationship but says nothing about minimal competence and the justification for a specific cutoff score. (*See id*.)

44. The expectancy analysis is more noteworthy.[30] Expectancy tables are intended to show the likelihood of a job candidate's attaining a defined level of job performance as a function of his or her predictor test scores. (Tr. Vol. 2, 513:17-514:10; Ex. 205 at p. 38.) Dr. Jeanneret's initial expert report sets forth expectancy tables based on the statistical relationship between Alert scores and ratings on the PDRF Composite for 190 incumbent Troopers in the validation sample.

45. Using the distribution of predicted PDRF Composite ratings of the Troopers in the sample, Dr. Jeanneret identified two alternative breakpoints to define "satisfactory" job performers: (a) those predicted to perform at or about the median[31] level of performance on the PDRF Composite (i.e., the top 50% of predicted performers, corresponding to a PDRF Composite rating of 198.02 or better); and (b) those predicted to perform at or above one standard deviation below the mean[32] of predicted performance on the PDRF Composite (approximately the top 85% of predicted

---

[30]At a couple of points in his testimony, Dr. Jeanneret described his expectancy analysis as a type of utility analysis. (*See* Tr.Vol. 2, 411:8–412:3; Tr. Vol. 5,1544:15-17.) Whether or not, on a theoretical plane, expectancy analysis is a subset or variety of utility analysis, Dr. Jeanneret's expectancy analysis is set out separately in his report in this case and was treated as a distinctly separate analysis in his testimony. It carries its own rationale, as compared to what was called the "utility analysis" in Dr. Jeanneret's report and testimony. Consequently, it is treated separately here.

[31]The median is the sample value having half the data above it and half the data below it. *See* Paetzold & Willborn at § 2.02, Table 2.1.

[32]The mean is the arithmetic average of sample values. *See* Paetzold & Wilborn at § 2.02, Table 2.1. The standard deviation is the square root of the variance, which is the sum of squared deviation around the mean. *Id*. Standard deviation is measured in the same units as the raw data. *Id*. In this case, Dr. Jeanneret testified that the standard deviation is 12.51 and the mean PDRF Composite score is 196.25. (Tr. Vol. 2, 536:16-537:7.) Therefore, one standard deviation below the mean is a PDRF Composite score of 183.74. (*Id*.)

performers, corresponding to a PDRF Composite rating of 183.74 or better). (Tr. Vol. 2, 522:15-523:5; 536:24-537:4; Ex. 205 at pp. 38-40.)

46. Defendants use Dr. Jeanneret's expectancy analysis to support their cutoff score of 75%, because, they say, it shows "100% of applicants selected at that cutoff would be expected to perform satisfactorily in the four dimensions of the job that comprise the PDRF Composite." (D.I. 301 at p. 21-22, ¶ 5.) Implicit in that assertion, of course, is that a lower Alert score would allow some into the Trooper ranks who would be less than satisfactory performers. That claim, however, ignores both the inherent imprecision of the expectancy analysis and the erroneous definition of "satisfactory" embedded in it.

47. The median level of predicted performance on the PDRF Composite (198.02) falls in Level 4, the category between "Outstanding" and "Expected" performance. (Tr. Vol. 2, 528:6-9.) Performance at one standard deviation below the mean (183.74) falls at the upper boundary of the "Expected" level. (Tr. Vol. 2, 537:13-20; Tr. Ex. 211.) Thus, remarkably, in Dr. Jeanneret's expectancy analysis, Trooper incumbents predicted to perform in the "Expected" level and even in the level above "Expected" would be characterized as performing less than satisfactorily. (*See* Tr. Vol. 2, 528:10-529:18; 537:13-538:2.)

48. Dr. Jeanneret conceded at trial, as well he should have, that "there's concern that that doesn't fully comply with the [*Lanning*] standard[,]" i.e., the standard of minimal competence. (Tr. Vol. 2, 411:5-412:2.)

49. My object, of course, is to fully comply with the *Lanning* standard, to determine whether the minimum level of literacy necessary to perform the job of

Trooper can be reflected in an Alert score and then to determine whether the score selected by the Defendants in fact reflects that minimum level of literacy.[33] Dr. Jeanneret's expectancy analysis is useful in that effort only to this extent: though it overstates the cutoff score required to reflect minimal competence on the job, by its own somewhat inflated terms it shows that 92.3% of applicants selected at a 70% Alert cutoff score would meet expectations. (Ex. 205 at p. 39). And, as is discussed more fully herein, *infra* at ¶¶ 65, 69-70, 85-86 and n. 42, to say that 92.3% will meet expectations is not to say that 7.7% will fall below the minimum qualifications for the job, both because "meet expectations" and "minimum qualifications" are not necessarily synonymous and because there is inevitably less certainty in these numbers than the precision of decimal points and percentages implies. Hence, the expectancy analysis undermines the Defendants' assertion that a 75% cutoff score on the Alert corresponds to minimum competence.

---

[33]In that regard, I bear in mind that while a valid test need not measure the totality of the skill set necessary for a job, the lack of representativeness (*see supra at* n.9) does reemphasize the importance of keeping the cutoff score at the minimum required level. (*See* Tr. Vol. 5, 1356:1-13.) That the cutoff score on a non-representative test must be set at the minimum level of the skills measured that are necessary to do the job is not an arbitrary legal standard. It is a standard with a psychometric and statistical rationale. Specifically, the use of a non-representative test with a cutoff score that exceeds the minimum may be counterproductive to the organizational goal of hiring the best overall performers, because an unnecessarily high cutoff score may well eliminate applicants who would be better overall performers on account of their strengths in other job-relevant areas. (Tr. Vol. 5, 1338:8-1340:23; Tr. Vol. 4, 1135:10-1145:16.) This is not a balancing of literacy against other required skills; it is, rather, a recognition that choosing an excessively high Alert cutoff score is counterproductive from both a public safety and a broader public policy perspective.

2. Dr. Wollack's two-step analysis

50. Dr. Wollack also sought to answer the question about the appropriate cutoff score on the Alert. He undertook an analysis which the parties came to refer to with the shorthand label, "the two-step analysis" or "two-step study." (Tr. Vol. 1, 96:19-24.) Dr. Wollack's two-step analysis consisted of the following: first, he asked supervisors what percentage of the officers under their charge had deficient reading and writing skills, and, second, he calculated an Alert cutoff that would eliminate that same percentage of applicants. (Tr. Vol. 1, 97:3-105:18.) Before undertaking that analysis for use in this case, Dr. Wollack had never conducted an Alert cutoff score analysis for the DSP Trooper job. (Tr. Vol. 1, 206:8-13; 92:16-19.) In setting cut-off scores on the Alert, the Defendants had followed general recommendations (Ex. 224 at p. 51; Tr. Vol. 3, 732:15-733:1) set forth in Dr. Wollack's publications. (Tr. Vol. 1, 210:11-16; 229:4-8; 92:20-93:12; Ex. 35, 37, 38, 39 and 40.) In November 1986, Dr. Wollack recommended that the Alert cutoff score be set at a raw score of 100 out of 160 (62.5%), regardless of test form. (Tr. Vol. 1, 208:13-209:1.) That recommendation was based on normative studies. (Tr. Vol. 1, 208:13-209:9; 212:5-20; Ex. 35 at p. 13.)

51. In November 1992, Dr. Wollack raised his recommended Alert cutoff to a raw score of 123 to 125 out of 160 (76.8 to 78.1%), regardless of test form.[34] (Tr. Vol.

[34]Later, in October 1995, Dr. Wollack again adjusted his recommendations for cutoff scores, this time based upon his recommended use of stanines to equate the difficulty level of the various forms of Alert. A stanine (a word coined by combining the words "standard" and "nine," *see Webster's Third New Internat'l Dictionary* at 2225 (1986)) is a portion of the area under a standard distribution curve. Stanines are created by dividing the area under the curve into nine intervals and numbering them from left right, with each interval being half of a standard deviation, so that the 5th stanine covers the midpoint or average portion of the curve, and stanines 1 through 4

1, 210:19-212:4; 213:10-19; 216:9-217:2; Ex. 37 at p. 20.) When asked at trial to explain this relatively dramatic increase in the recommended cutoff score, Dr. Wollack stated that two studies his company had conducted in 1990 and 1991, in Texas and Washington, respectively, had provided the first opportunity for him to ascertain what "job-related cutoffs" should be. (Tr. Vol. 1, 211:1-8.) He said, "[i]t wasn't until 1990 that we did our first of the two-step studies in which we related the scores on the Alert examination to job performance. And when we did that, we realized that the cutoff score recommendations that we had been making ... were way too low." (Tr. Vol. 1, 211:13-19.)

52. As part of the two-step analysis in the 1990 Texas study, supervisors were asked to estimate the percentage of police officers whom they had supervised during the prior five-year period who had deficient reading and writing skills. The average of the supervisors' estimates was 17.6%. (Tr. Vol. 1, 214:2-5.) Dr. Wollack administered the Alert to a sample of incumbent police officers in Texas and determined that an Alert cutoff score of 123 (76.8%) would have prevented the hire of 17.6% of the incumbent sample. (Tr. Vol. 1, 214:11-20.) In the 1991 Washington study, Dr. Wollack followed the same procedure. The supervisors in Washington returned an average estimate of

---

represent below average and stanines 6 through 9 represent above average. Before his 1995 recommendation, Dr. Wollack's Alert cutoff score recommendations had not accounted for the variation in difficulty between Alert forms. Dr. Wollack's new position called for the Alert cutoff score to be set at the lower bound of stanine 5, because stanine 5 for the most part encompassed the 123 to 125 raw score range. (Tr. Vol. 1, 231:19-23; 232:15-19; 235:12-21; Tr. Ex. 39 at p. 2; Tr. Ex. 40 at p. 2.) An exception to the foregoing statement had to do with Alert Form 07. At the time Defendants used Alert Form 07, Dr. Wollack had not yet published a cutoff score recommendation for it, but Defendants based their cutoff score on advice from Dr. Wollack. (D.I. 263 at p. 5, ¶ 27; Tr. Vol. 1, 235:12-21; Tr. Ex. 40 at 2.)

12% and Dr. Wollack determined that an Alert cutoff score of 125 (78.1%) would have prevented the hire of 12% of the incumbent sample. (Tr. Vol. 1, 214:21-215:24.)[35]

53. In his two-step studies, Dr. Wollack did not remove outliers before averaging the supervisors' estimates, nor did he take any steps to corroborate those estimates. (Tr. Vol. 1, 214:7-10; 215:5-17; 218:1-20; 233:18-21.) Dr. Wollack assumed that his normative samples would have the same percentage of individuals with deficient reading and writing skills as the populations on which the supervisors' estimates were based, even though they were different groups of people. (*See* Tr. Vol. 1, 220:3-20.) Dr. Wollack further assumed that incumbents with deficient reading and writing skills would necessarily obtain the lowest Alert scores during normative testing, but he did nothing to determine whether the individuals who would be eliminated by his recommended cutoff scores in fact had deficient reading and writing skills. (Tr. Vol. 1, 219:6-226:2; 226:22-227:8; 227:14-23.)

54. As part of his study for this case, Dr. Wollack used his two-step analysis to assess whether Defendants' Alert cutoff scores corresponded to the minimum level of reading and writing skills necessary for successful job performance. (Tr. Vol. 1, 236:3-8; Ex. 224 at p. 51.) Dr. Wollack asked DSP supervisors to estimate what percentage of Troopers they had supervised over an eight-year period (January 1992 through

---

[35]In 1997, Dr. Wollack conducted a two-step cutoff score analysis with the Missouri State Highway Patrol. The supervisors returned an average estimate of 3.9% and Dr. Wollack determined that an Alert cutoff score of 124 (77.5%) would eliminate the bottom 3.9% of the incumbent sample to whom he administered the Alert. Dr. Wollack's two-step analysis in Missouri did not change his published Alert cutoff score recommendation of 123-125, because it yielded a cutoff score within that range. (Tr. Vol. 1, 233:1-234:2.)

December 1999) had unsatisfactory reading and writing skills. (Tr. Vol. 1, 236:16-20.) The supervisors returned an average estimate of 4.58%. (Tr. Vol. 1, 236:21-23.) Dr. Wollack then applied the supervisors' estimates to the raw Alert scores obtained at the time of selection by the 269 Troopers hired during the seven-year period at issue in this case (1992-1998), and determined that an Alert cutoff score of 122 (76.2%) would have eliminated the lowest 4.58% of the Alert score distribution. (Tr. Vol. 1, 236:24-238:7; 239:3-9; Ex. 224 at Appendix N.)

55. The DSP supervisors were not provided with a list of the Troopers they supervised during the eight-year period for which they were asked to provide an estimate. (Tr. Vol. 1; 239:10-14.) The individuals supervised during that period included individuals hired before 1992. Although Defendants used the Alert starting in 1981, there is no evidence regarding the Alert cutoff scores used before the 1992 Alert administrations. (Tr. Vol. 1, 240:17-241:1.) However, Dr. Wollack in 1986 recommended an Alert cutoff score of 100 out of 160 (62.5%), and he did not raise his Alert cutoff score recommendations until after he conducted his two-step studies in 1990 and 1991. (Tr. Vol. 1, 241:12-20.)

56. Based on the results from his two-step analysis in the DSP, Dr. Wollack concluded that an Alert cutoff score of 122 (76.2%) is appropriate because that cutoff score would have prevented 4.58% of those hired between 1992 and 1998 from being further considered for hire, even though the supervisors' 4.58% estimate may relate to individuals hired during a different time period (Tr. Vol. 1, 245:29-246:12), and despite the fact that no applicant hired as a DSP Trooper during the period in question was

terminated for sub-standard reading or writing skills, nor did any such individual resign in lieu of or in anticipation of being fired for those reasons. (D.I. 263 at 6, ¶ 32.)

57. Dr. Wollack did nothing to corroborate his assumption that the individuals with the lowest Alert scores perform the worst on the reading and writing aspects of their job. (Tr. Vol. 263:22-264:9.) Dr. Wollack never collected or examined any information about the job performance of the incumbents with the lowest Alert scores. (*Id*.) In fact, eighteen of the 55 DSP supervisors in the study estimated that zero percent of the Troopers they had supervised had unsatisfactory reading and writing skills. (Tr. Vol. 1, 263:6-11.) Dr. Wollack admitted that those eighteen supervisors may have supervised the Troopers with the lowest Alert scores. (Tr. Vol. 1, 263:12-21.)

58. Of course, as Dr. Wollack admitted, the two-step analysis he has repeatedly followed is guaranteed to result in a recommended cutoff score equal to or higher than the Alert cutoff score used by the police department in hiring the job incumbents being studied. (*See* Tr. Vol. 1, 113:21-24; 269:11-17.) As he put it, "the cutoff that you derive from this process cannot be lower than the lowest score of the incumbents in the group." (Tr. Vol. 1, 113:21-24.) Dr. Wollack's approach also incorrectly assumes a perfect correlation between Alert scores and performance in the reading and writing aspects of a Trooper's job.[36] Significantly, his approach does not, and cannot, consider the Alert scores of those who never passed the test but who, in fact, might have the reading and writing skills necessary to do the job. (Tr. Vol. 1,

[36]The Defendants' own statistical evidence is to the contrary. It shows only modest correlations between Alert scores and the literacy demands of the job, as reflected in the PDRF Composite. (*See supra* at ¶ 39.)

269:18-271:2.) It cannot consider them because, by definition, those candidates were never hired. The two-step analysis thus assumes the answer it is trying to prove, namely, that a failing score on the Alert means sub-minimal reading and writing skills for the job of Trooper. It is, in short, an elaborate exercise in question-begging and entirely unpersuasive on the central question before me.

59. Dr. Wollack's cutoff score conclusions in this case are all the more puzzling because many jurisdictions use the Alert with lower cutoff scores than those used by the Defendants. (Tr. Vol. 1, 271:3-8.) Dr. Wollack does not disagree with the use of those lower scores (Tr. Vol. 1, 272:8-17; 273:8-11), even though he believes that police officer jobs throughout the country are highly similar and the required reading and writing skills are the same for virtually every law enforcement agency (Tr. Vol. 1, 95:4-11; 544:9-14).[37] In that same vein, he previously recommended a significantly lower Alert cutoff score (*see supra* at ¶ 50) and provided no evidence to demonstrate that the individuals who became police officers at his lower recommended cutoff score possessed inadequate reading and writing skills.

60. Finally, and not insignificantly, Dr. Wollack's conclusion about an appropriate cutoff score is undermined by his acknowledgment that the standard error of measurement on the Alert[38] is such that Alert scores differing by as much as 6.5

---

[37]Indeed, as previously noted (*supra* at ¶ 7), the parties have stipulated that the reading and writing demands of the entry-level law enforcement job are essentially the same from jurisdiction to jurisdiction. (D.I. 263 at p. 5, ¶ 26.)

[38]The "standard error of measurement" of a statistic is the standard deviation of the sampling distribution of that statistic. The term "statistic" itself is defined "as a value computed from a sample." W. Curtis, *Statistical Concepts for Attorneys: A Reference Guide* at 97 (1983). From any population, a variety of samples may be drawn, with the

points may not represent any difference in skill level. (*See* Ex. 225 at 9-10; Tr. 282:18 – 283:21.) Thus, for example, a Trooper candidate who scored 111 on Form 06 of the Alert, which is a score under 70%, cannot be meaningfully differentiated from someone with a passing score of 117, or approximately 73%, on that Form. (*See id.*)

3. False positives and false negatives

61. Both sides in this dispute have invested significant time in arguing about evidence of "false positives" and "false negatives" in the testing results from the Defendants' use of the Alert. A false positive in this context means a job candidate who took the Alert and passed but who actually had sub-minimal literacy skills. Conversely, a false negative is a candidate who failed the Alert but who in fact had at least the minimal literacy skills for the job.

a. False positives

62. Dr. Jeanneret observed that, of the 190 Troopers in the validation sample, 13 or 14 individuals had ratings below 144.54 on the PDRF Composite.[39] He then opined that these individuals were false positives and that lowering the Alert cutoff score

---

same statistic computed from each sample having a different value. *Id.* at 97-98. One could take the various samples and construct a sampling distribution having its own mean and variance. That then becomes the basis for calculating the standard error of measurement, as "[t]he square root of the variance of a sampling distribution is called the 'standard error' to distinguish it from the standard deviation of the population." *Id.* at 98. The standard error is useful because it shows the amount of fluctuation in a statistic.

[39]According to Dr. Jeanneret, the ambiguity in the number is due to rounding. However, Ex. 32 shows 13 Trooper incumbents with a PDRF Composite rating below 144.54. (Ex. 32 at pp. 4-6.)

below 75% would result in the hiring of additional Troopers who lack the necessary literacy skills. However, I did not see any persuasive evidence to support that assertion.

63. First, as noted earlier (*supra* at ¶ 37), Dr. Jeanneret's PDRF rating form did not identify a score or range of scores that represents the minimum level of acceptable performance on a given job dimension. (Tr. Vol. 2, 599:7-16.) Dr. Jeanneret acknowledged that his rating scale did not use the term "minimum acceptable performance," that the supervisors were not advised as to what point on the rating scale corresponds to minimally acceptable performance, and that the supervisors were not asked to make such a judgment. (Tr. Vol. 2, 599:17-600:5; 602:14-604:18.)

64. In fact, the supervisors were asked to rate a Trooper's exhibited performance on job dimensions such as written communication and oral communication, but not whether the Trooper had deficient skills in these areas. (Tr. Vol. 3, 617:7-618:4.) Even if I were inclined to equate the rating of "Expected" with minimum acceptable skill level, there is precious little evidence to justify that step. Because supervisors were not asked to provide the reasons why an individual's performance was considered to be below "Expected" (Tr. Vol. 3, 617:12-16; 623:13-16), there was no evidence that the lowest-ranked Troopers fail to perform as expected because they lack necessary skills rather than because of some other equally plausible reason, such as attitude or motivation problems. Dr. Jeanneret did not conduct any further study of the reading and writing skills of the 13 individuals rated below 144.54 on the PDRF Composite (Tr. Vol. 3, 623:19-23), and Defendants called no witnesses with personal knowledge of the job performance of the alleged false positives.

65. The information the Defendants provided in support of their "false positive" argument actually leads to conclusions contrary to their position. Of the 13 Troopers in the validation sample who were rated below 144.54 on the PDRF Composite, only one received a PDRF Composite rating in the "Poor" level; the other 12 individuals received PDRF Composite ratings in the unlabeled category between "Poor" and "Expected." (Tr. Vol. 2, 595:13-15; Ex. 32 at pp. 4-6.) Several of the 13 individuals received PDRF Composite ratings just below the borderline of the "Expected" level (*e.g.*, 143.26, 141.77, 141.75, 141.45). (Ex. 32 at pp. 4-6.) And, significantly, some of them scored relatively high on the Alert (*e.g.*, 86.88%, 86.25%, 85.0%). (Ex. 32 at pp. 4-6.) In fact, the Trooper in the validation sample with the highest Alert score (152 items out of 160 correct, or 95%) narrowly missed being rated below the "Expected" level on the PDRF Composite (148.33). (Tr. Vol. 2, 597:22-599:6; Ex. 32 at p. 5.) Those facts suggest that a score of 144.54 on the PDRF Composite does not equate to a lack of the minimum literacy skills for the job of Trooper. They also serve to emphasize the attenuated predictive power of the Alert.[40]

b. False negatives

66. The United States submitted a list of 97 individuals who failed the Alert but who completed law enforcement training in other jurisdictions and obtained law enforcement certification and employment. (Ex. 10.) The United States argued that these 97 individuals are "false negatives," in other words, they are candidates who were

[40]Dr. Siskin's rebuttal report to Dr. Jeanneret's expert report (Ex. 8) also highlights this point. Dr. Siskin states that "[t]he true differences in expected performance on the PDRF Composite between the pass groups that would result from the use of the various cutoffs Dr. Jeanneret considered are quite trivial." (Ex. 8 at p. 9.)

screened out by the Alert but who in reality had the minimal literacy skills for the Trooper job. Because the 97 candidates in question obtained law enforcement employment, the United States argues that they must have at least the minimum reading and writing skills, as it is undisputed that the reading and writing demands of the entry-level law enforcement job are the same across jurisdictions. (Tr. Vol. 1, 95:4-11; 54:9-14; D.I. 263 at pp. 5-6, ¶¶ 26 and 34.)

67. While the "false negatives" evidence is not beyond question,[41] I found it persuasive, particularly as to those failing Trooper candidates who joined other police organizations in Delaware. The same academy training is provided in combined classes to new DSP recruits and to recruits from local law enforcement agencies. (Tr. Vol. 3, 700:2-6.) The DSP and local recruits are trained side-by-side in the same classrooms with the same instructors, course materials, and tests, and the reading and

[41]Drs. Wollack and Jeanneret testified that there are numerous factors that should have been considered in compiling the list of false negatives, but which were not. Specifically, the United States did not take into account any of the non-test-related reasons why the 97 individuals might have failed Alert, such as poor health, stress, inadequate sleep, or the testing environment. (Tr. Vol. 4, 1203:9-12; Tr. Vol. 5, 1453:1-10; Exs. 8 & 24.) Dr. Goldstein agreed that there are, in fact, non-test-related factors that could explain a failing test score but which are irrelevant to the appropriateness of the cutoff on that test. (Tr. Vol. 5, 1460:10-17.) The United States also did not consider the time that elapsed between the date on which any of the 97 individuals failed the Alert at the DSP and the dates they graduated from an academy, were certified, and hired. (Tr. Vol. 4, 1214:24-1215:7.) Dr. Wollack pointed out that numerous individuals on the United States' list failed the Alert several years before they eventually obtained certification of employment elsewhere. (Tr. Vol. Vol. 1, 140:10-141:24; Ex. 225 at pp. 6-7.) Dr. Goldstein agreed that intervening learning might occur during that time, and that such learning would be relevant in considering whether an individual is truly a false negative. (Tr. Vol. 5, 1460:2-9.) In fact, one of the witnesses presented by the United States as a "false negative" acknowledged that he had taken steps to improve his reading and writing skills after he failed the Alert at the DSP but before he obtained a job elsewhere. (Tr. Vol. 4, 919:6-921:10.)

writing skills required to complete the training academy are the same for DSP and local recruits. (Tr. Vol. 3, 700:2-701:1; Ex. 138, 18:13-23:4.) It is therefore noteworthy that the local recruits generally performed as well on the academy tests as the DSP recruits. (*See* Ex. 152, summarizing data from Ex. 121-129.)

68. Eleven of the 97 individuals identified as false negatives testified at trial. These eleven individuals were employed by various law enforcement agencies, including the New Castle County Police Department, the Delaware Division of Alcohol and Tobacco Enforcement, the Camden New Jersey Police Department, the Philadelphia Police Department, the United States Secret Service, the Salisbury Maryland Policy Department, the Freehold New Jersey Police Department, and the Wilmington Police Department. Each of these individuals testified that he was able to perform the reading and writing tasks of the law enforcement job. Many had been promoted and had received commendations. Several held Bachelor's degrees at the time they took and failed the Alert when applying to join the DSP. (Tr. Vol. 4, 909:18-1014:19.)

69. The United States, of course, would like me to conclude from the testimony of those eleven officers that the other 86 on their false negatives list are similarly successful in their law enforcement careers. While there is no direct evidence in the record to support that assumption, there is at least a fair inference to be drawn that some additional and significant number of officers[42] who failed the Alert are

[42]The Defendants argue that the list of false negatives is inflated. Among other things, they assert that the list includes many applicants who barely failed the Alert when they applied to the DSP and whose scores fall within a standard error of measurement. They claim that those individuals should not be viewed as false

currently employed in the law enforcement field in other jurisdictions and are performing with at least the requisite level of skill to maintain their positions. Unless one is to presume that the departments they work for are keeping them on staff despite incompetence, a cynical conclusion for which there is no evidence,[43] the most logical conclusion is that those officers had the minimal literacy skills to do the Trooper job[44] but were falsely screened out of consideration.

---

negatives. (*See* D.I. 301 at ¶ 97; Tr. Vol. 1, 137:7-138:16; 502:18-504:20.) This head-in-the-sand approach cannot withstand logical review. An individual who is screened out of consideration by the Alert but who does in fact have the literacy skills for the Trooper job cannot in fairness be viewed as anything but a falsely identified member of a negative (i.e., "unqualified") category, notwithstanding the Defendants' after-the-fact rationalization about statistical concepts like standard error of measurement. Of course, the existence of false negatives does not make the Alert an inappropriate screening mechanism. Perfection is an impossible goal in employment selection as it is in virtually all human endeavors, but achieving the legally required goal of identifying the minimum skill level required for a particular job is not assisted by ignoring or rationalizing away evidence of false negatives.

[43]The Defendants assert that various administrative hurdles stand in the way of terminating a Trooper, such as formal grievance procedures, union involvement, the convening of a trial board, hearings, and appeals. (D.I. 301 at ¶ 21.) While the termination process within the DSP and on other police forces no doubt involves similar transaction costs, no one presented evidence that officers lacking basic competence and literacy remain employed for any significant length of time. This is not surprising, given that such an admission by the State would necessarily mean there has been supervisory incompetence of a more serious nature than literacy incompetence among entry level Troopers. For the same reason that I would not, without persuasive evidence to the contrary, believe that DSP supervisors are failing in their responsibility to maintain a competent police force, I will not assume that police supervisors in other jurisdictions are keeping incompetent officers on the job.

[44]Again, the parties agree that the reading and writing demands on entry level law enforcement officers, including DSP Troopers, are basically the same throughout the United States. (*See* D.I. 263 at p. 5, ¶ 26.) I recognize that it is possible for a candidate's literacy skills to have improved from the time he or she failed the Alert, but while that possibility lessens the impact of the false negatives evidence, it does not eliminate it.

70. It is particularly noteworthy that, among the 97 individuals on the false negatives list, two-thirds of them scored approximately 70% on the Alert.[45] (Tr. Vol. 2, 503:23-504:10; Ex. 10.) That fact undermines the Defendants' position that 75% was an appropriate cutoff score on the Alert, but it also shows that the evidence regarding false negatives does not support the United States' contention that a more appropriate Alert cutoff score was 60%.

4. Regression analysis

71. Both sides presented evidence regarding linear regression analysis conducted on the Alert scores and PDRF Composite information collected in this case. Linear regression is an analytical technique that examines the relationship between two variables by plotting data on the X (horizontal) and Y (vertical) axes of a graph and then determining the line that best fits through those data points by minimizing the distance between each point and the line itself. The resulting line is known as the "least squares regression line." The regression line has a defined slope and an intercept value that indicates the point at which it crosses the Y axis. (*See* Tr. Vol. 2, 426:7-428:8.) Regression analysis is helpful in predicting an unknown value from a correlated known value.

---

[45]The standardized Alert scores of the 11 "false negative" applicants who testified are as follows:

1. 76.25	4. 73.75	7. 71.88	10. 68.75
2. 74.38	5. 73.13	8. 70.63	11. 57.50
3. 73.75	6. 73.13	9. 70.63	

(*See* Ex. 10.)

72. As with much of the evidence in this case, the parties have taken the same data, analyzed it with nominally objective, mathematical tools, and yet managed to reach dramatically different conclusions. With regard to regression analysis, the difference between the parties' conclusions hinges upon whether they chose a "forward regression" analysis or a "reverse regression" analysis.

73. The United States chose to treat the Alert scores as the known value and Trooper performance as the unknown value. Treating test scores as the known value and performance as the unknown, "to-be-predicted" value is a typical approach in assessing the validity of employment tests. In this case, the United States' approach is labeled "forward regression" to distinguish it from the Defendants' "reverse regression" analysis of the data. I will first address the Defendants' analysis.

a. Reverse regression

74. The Defendants argue that setting a cutoff score is a different matter than establishing test validity and that it therefore requires a different approach. Their more novel[46] reverse regression approach treats performance as the known value and Alert scores as the to-be-predicted value. That approach, they say, is more appropriate for determining what Alert score corresponds to minimally acceptable performance in the

[46]Prior to this case, Dr. Jeanneret had never used the reverse regression method (Tr. Vol. 2, 563:13-21), and no evidence was presented that it has ever been used to analyze a cutoff score in a case such as this. The journal article on which the approach is based includes the following cautionary advice: "One point bears reemphasis. No research has yet appeared on any of these proposed methods. Use of any of them will require a thorough assessment of reliability until a body of research literature develops." (Ex. 117 at p. 19.)

literacy dimension of a DSP Trooper's job, since they claim to have captured the minimally acceptable performance level in a specific PDRF Composite score.

75. Interestingly, one of the United States' experts, Dr. Goldstein, is the one who initially suggested using the reverse regression approach in the present case. (*See* Tr. Vol. 5, 1463:12-15.) In his expert report (Ex. 24 at p. 13), Dr. Goldstein quoted from a professional article published in *Personnel Psychology* in 1988, entitled "Setting Cutoff Scores: Legal, Psychometric and Professional Issues and Guidelines," by Drs. Wayne Cascio, Ralph Alexander, and Gerald Barrett. (Ex. 117; the "Cascio article".) The methodology laid out in the Cascio article, referred to therein as "Research Suggestion No. 7", is the reverse regression approach adopted by the Defendants and involves the regression of test scores on to job performance.[47] In theory, one can take a known minimum performance level and, using regression, predict the specific Alert score associated with that performance level. (Tr. Vol. 2, 429:24-432:21; Ex. 117 at p. 17.) Dr. Jeanneret followed through on Dr. Goldstein's suggestion, using 144.5 on the PDRF Composite as the minimally acceptable level of literacy performance, and then seeking as the unknown value the Alert score associated with that PDRF Composite score. (Tr. Vol. 2, 432:1-21.) Dr. Jeanneret confirmed the propriety of that analytical approach with Dr. Cascio, the author of the article that Dr. Goldstein had cited. (Tr. Vol. 2, 454:15-455:21.)

[47]Dr. Goldstein retreated from his suggestion at trial, saying it was a "mistake," (Tr. Vol. 5, 1465:18) but acknowledged that he had previously suggested it as the correct approach for setting the cutoff score. (Tr. Vol. 5, 1464:22 – 1467:2.)

76. When Dr. Jeanneret undertook the reverse regression analysis on the data in this case, it at first indicated that an Alert score of 81% is the score that corresponds to 144.5 on the PDRF Composite. (Tr. Vol. 2, 455:22-456:12; Ex. 298.) However, as Dr. Jeanneret admitted at trial, the data set he was working from had a significant problem: with the limited exception of those few Troopers who had failed the Alert but passed it at a later date,[48] the data did not include, because it obviously could not, data on the performance of test-takers who failed the Alert, since they were never hired as Troopers. (Tr. Vol. 2, 459:2-460:13.) That lack of pertinent data creates what has been referred to variously as a "restriction in range" problem, or a "limited dependent variable" problem, or a "truncated distribution" problem.[49] (*Id.*)

77. As a consequence of the truncated distribution problem, Dr. Jeanneret's reverse regression model is rendered meaningless, without some kind of correction. As Dr. Jeanneret conceded, it is mathematically impossible to predict an Alert cutoff score below the cutoff score used by Defendants using the reverse regression method on the data set in this case. (Tr. Vol. 3, 631:4-8.) That is because the constant, or y-intercept, in Dr. Jeanneret's equation is an Alert score of 75.4%, which is above the standardized Alert cutoff score of 75% used by the DSP. (Tr. Vol. 3, 626:10-627:11; Tr. Ex. 298.)

---

[48]Dr. Jeanneret estimated that 10% of the data represented Troopers who had failed the Alert at least once. (Tr. Vol. 2, 459:23 – 460:3.) More precisely, there are 18 individuals who scored below a standardized score of 75% on their first Alert administration, but were subsequently hired based on a later Alert score or by passing Defendants' replacement test. (Tr. Vol. 2, 579:11-24; Ex. 32.)

[49]Dr. Goldstein did admit that a reverse regression analysis would be "optimal...with a full set of data[,]" but that was not present in this case. (Tr. Vol. 4, 1076:15-19.)

Thus, using any PDRF Composite score above zero in the regression equation will result in a predicted Alert score above the cutoff score used by the DSP.[50] Dr. Jeanneret conceded that, using this regression line, even the worst possible performer on the PDRF Composite (an individual with a rating of one on the 1-60 scale on each of the four dimensions that comprise the PDRF Composite) is predicted to obtain a passing score of 77.7% on the Alert. (Tr. Vol. 3, 632:10-633:22.) Dr. Siskin computed that a PDRF Composite score of negative 13 would be required to yield a predicted cutoff score below the cutoff score actually used by the DPS. (Tr. Vol. 4, 1093:22-1094:23.) Thus, use of the reverse regression method in this context is mathematically guaranteed to arrive at a result favorable to Defendants.

78. Furthermore, the 81% Alert cutoff score determined by the reverse regression method produces the incongruous result that numerous Trooper incumbents rated by their supervisors as performing at the "Expected" level or better on the PDRF Composite would have failed the Alert if the cutoff score had been 81%. Of the 190 incumbents, a total of 62 scored below 81% on their first Alert administration; of those 62 incumbents, 10 were rated "Outstanding" on the PDRF Composite; 27 were rated between "Outstanding" and "Expected"; and 16 were rated at the "Expected" level. (Tr. Vol. 2, 569:14-570:22; Tr. Ex. 32.) The fact that large numbers of incumbent Troopers

[50] The Defendants dispute this point and claim that "in the cited portion of the record, Dr. Jeanneret testified that it is the 'limited dependent variable problem or the fact that we have restriction in range,' plus the fact that 'as long as the PDRF score is not a negative number, the predicted Alert score is going to be above 75.4 percent, that leads to the predicted cutoff score of 75%.'" (D.I. 303 at 8.) The fact remains, however, that the uncorrected data range ensures that the predicted cutoff score will be above 75%, making the analysis tantamount to a self-fulfilling prophecy.

who scored below 81% are performing successfully on the job dimensions correlated with the Alert is conclusive evidence that an Alert cutoff score of 81% does not correspond to the minimum level of skills necessary to perform the job.

79. Dr. Jeanneret attempted to correct for the truncated distribution problem and asserted that, following his correction,[51] the predicted Alert cutoff score that corresponds to minimally acceptable literacy performance by a Trooper "drops down to about 72 percent, 73 percent, maybe 75 percent correct." (Tr. Vol. 2, 460:18-19.)

80. I find that Dr. Jeanneret's reverse regression result, even after his attempted correction for the restriction in range, is less than persuasive, because of the problems already noted. (*See supra* at ¶¶ 76-78.) However, just because I am not persuaded of the conclusion that Dr. Jeanneret has advanced to justify the Defendants' cutoff score, that does not mean that the reverse regression approach is devoid of evidentiary value. It does deserve further consideration because both sides have acknowledged ways in which, in theory at least, the reverse regression approach can shed light on the question of minimally acceptable literacy performance.

81. The plaintiffs, through Dr. Siskin, argue that the reverse regression line must be corrected to account for the conditional distribution of data around the cutoff point on the regression line. I agree. Except in the case of perfectly correlated variables, every regression line, has, by definition, some distribution of data points

[51]The correction itself presents logical difficulties. Dr. Jeanneret's corrections were not in his report and were first mentioned at trial. (Tr. Vol. 3, 631:9-632:9.) Accordingly, Dr. Siskin testified that he could not comment on the assumptions inherent in Dr. Jeanneret's simulations. (Tr. Vol. 4, 1077:13-19.) However, Dr. Siskin noted that, based on his review of the research literature, "nobody has been able to solve the problem about how you handle [a] truncated database." (Tr. Vol. 4, 1076:23-1079:1.)

around it. A regression line represents the best estimate of the value of the dependent variable (y) for a given independent variable (x); that is, for a given value of x, the corresponding point on the regression line can be considered the average, or mean, of the potential y-values. (Tr. Vol. 4, 1036:19-1038:1.) In theory, for each given value of x on the regression line, there is a symmetrical distribution curve of potential y-values around it, with 50% of the values above and 50% below the regression line. (Tr. Vol. 4, 1038:2-1039:17; 1064:10-1067:5; Tr. Vol. 5, 1532:14-1433:20.)

82. The size of the correlation coefficient is an indication of the degree of variance around the regression line. (Tr. Vol. 4, 1040:2-1042:10.) A perfect correlation of 1 means that every data point falls on the regression line. The lower the correlation coefficient, the greater the variation of points around the line. (Tr. Vol. 4, 1066:14-1067:5.) Low correlations are associated with more error in prediction. Because the regression line represents the mean of potential y-values for a given value of x, greater variation around the mean translates into more prediction error. (Tr. Vol. 4, 1065:3-1067:5.)

83. In the present case, where the correlation between the Alert and PDRF Composite variables generally falls into the low range, there is a substantial amount of variance around the regression line and consequently, there is a greater potential for errors in prediction. This point is illustrated graphically by the scatterplot of observed Alert and PDRF Composite values of the 190 incumbents in the validation sample. (Ex. 301.)

84. Using his reverse regression analysis, Dr. Jeanneret identified an Alert cutoff score by locating the point on the regression line that intersects with a PDRF

Composite value of 144.5, a point that is somewhere in the low to mid 70th percentile, assuming that Dr. Jeanneret's corrections for the truncated distribution are accurate. (Tr. Vol. 2, 460:18-19.) However, because the regression line represents the mean (*see supra* ¶ 81) one-half of the individuals performing at a level of 144.5 on the PDRF Composite may theoretically have Alert scores below that selected cutoff point. As Dr. Siskin pointed out, given Dr. Jeanneret's reverse regression results, 50% of applicants who would perform at the Expected level of 144.5 on the PDRF Composite would be predicted to fail the Alert using Dr. Jeanneret's cutoff score.[52] (Tr. Vol. 4, 1079:4-1080:3; 1096:10-1097:6; Tr. Vol. 5, 1533:11-15349.)

85. The validation sample of 190 incumbent Troopers demonstrates this point. There are 18 individuals in the validation sample who scored below a standardized score of 75% on their first Alert administration, but who subsequently re-tested and were hired by the DSP. (Tr. Vol. 2, 579:22-24; Ex. 32.) Of these 18 individuals, 14 were rated as performing successfully on the PDRF Composite job dimensions: one was rated "Outstanding"; eight were rated between "Outstanding" and "Expected"; and five were rated as performing at the "Expected" level. (Tr. Vol. 2, 580:1-19; Ex. 32; Ex. 301.) The Trooper who was rated "Outstanding" on the PDRF Composite failed the

[52]Dr. Siskin further testified that approximately one-third of applicants who would perform at a PDRF Composite level of 200, which falls in the level above "Expected", would be predicted to fail the Alert using Dr. Jeanneret's cutoff score. Approximately one-fifth of applicants who would perform at a PDRF Composite level of 270, in the "Outstanding" range, would be predicted to fail the Alert. (Tr. Vol. 4, 1097:7-1098:2.)

Alert three times before passing it on the fourth attempt. (Tr. Vol. 2, 580:20-582:12; Ex. 32; Ex. 262.)[53]

86. By recommending that the cutoff score be set at the point on the reverse regression line that intersects with a PDRF Composite value of 144.5, even after correcting for the truncated distribution data, Dr. Jeanneret recommends a cutoff score that will result in erroneous predictions half of the time. (Tr. Vol. 4, 1107:6-20.) Indeed, he conceded that if his recommended cutoff score were used, 50% of the applicants who would perform at the "Expected" level of 144.5 on the PDRF Composite would fail the Alert. (Tr. Vol. 5, 1533:3-1534:9; *see also* Tr. Vol. 4, 1079:4-1080:3; 1096:10-1097:6.) That result is obviously unacceptable. A cutoff score that eliminates half the individuals who would perform at an acceptable level cannot be described as corresponding with the minimum skill level necessary to do the job.

87. Dr. Siskin described one method of using information about the conditional distribution of data around the reverse regression line (referred to as the "standard error of the regression") to identify a cutoff score that more closely corresponds to the minimum skill level necessary to perform the job. The standard error of the reverse regression line prepared by Dr. Jeanneret is 5.8 percentage points on the Alert. (Tr.

[53]As previously noted (*supra* at ¶ 56), no applicant hired as a Trooper during the time period relevant to this case was terminated for substandard reading or writing skills, nor did any such individual resign in lieu or in anticipation of being terminated for substandard reading or writing skills. (D.I. 263 at p. 6, ¶ 32.) Seven Troopers hired during the relevant period were terminated or resigned in lieu of termination for reasons unrelated to reading or writing skills. (Tr. Vol. 3, 769:3-22; Tr. Ex. 132; D.I. 302 at p. 4, ¶ 12.)

Vol. 4, 1080:8:12.) By multiplying this standard error by 1.645 standard deviations[54] and subtracting that value from the mean point on the regression line (*i.e.*, 75%), Dr. Siskin determined the Alert score that would include 95% of individuals who would perform the job at the Expected level of 144.5. Here, that is an Alert score slightly higher than 65% (*i.e.*, 75% minus (5.8% x 1.645) equals approximately 66%). (Tr. Vol. 4, 1079:4-1081:18; 1038:2-23; Ex. 160.) Thus, Dr. Jeanneret's analysis, if accepted despite its flaws, and after being adjusted to account for the conditional distribution of data, points to a cutoff score above 65%.

b. Forward regression

88. On behalf of the Plaintiff, Dr. Siskin undertook the regression of job performance, as reflected in the PDRF Composite, on to the Alert scores. The result, however was a wholly unreasonable cutoff score of 43%. (Tr. 1295:18-1296:6; Ex. 8 at p. 2.) Dr. Siskin himself stated that a 43% cutoff was unreasonable. (Tr. 1105:8-18.) At that cutoff score, an applicant would have a very high likelihood of failing to meet a minimum level of performance. Moreover, in order to accept the forward regression prediction, one must again struggle with a lack of data by assuming a linear relationship between Alert scores and PDRF Composite scores far beyond what has actually been observed. (Tr. Vol. 4,1105:20-1107:5; *see* Ex. 8 at 6 n.1.) Dr. Siskin testified that making that assumption would require pure guesswork, and he would never advise an

[54]Dr. Siskin selected 1.645 standard deviations below the mean because he "want[ed] to get to the five percent level," in other words, determine a score that at least 95% of test-takers would achieve. (Tr. Vol. 4, 1080:13-1081:4.) Thus, Dr. Siskin also engaged in a bit of question-begging too (*see supra* at ¶ 58) by selecting particular numbers in order to reach his desired result. Regardless, I do agree with Dr. Siskin that an acceptable cutoff score on the Alert cannot have a 50% failure rate.

employer to use that model to make such predictions. (Tr. Vol. 5, 1297:24-1299:16.) In fact, he would warn anyone against it. (Tr. Vol. 5, 1300:9-12.) The result of the forward regression analysis in this case is thus entirely unpersuasive.

5. The character of the Trooper job

89. Focusing on the character of the Trooper job, both parties offered non-statistical evidence to support their respective positions about the appropriate cutoff score in this case. For its part, the United States emphasizes what Dr. Goldstein described as the "routine" nature of the reading and writing demands on Troopers.

90. Dr. Goldstein testified that the Defendants' cutoff scores on the Alert were too high, given the nature of the reading and writing demands of the DSP Trooper job and the level of education attained by the applicant population. Dr. Goldstein further testified that a standardized Alert cutoff score of 60% would have been more appropriate. He stated his opinion that the reading and writing tasks of the Trooper job are not difficult for individuals with at least 60 hours of college credit. The reports written by Troopers are repetitive, he asserted, and many of the most commonly used report forms are short and simple. Dr. Goldstein testified that the most complex reports created by the Troopers have a narrative section that is typically only about a quarter to one half of a page in length. (Tr. Vol. 5, 1361:9-1363:17.)

91. The Defendants, naturally, view the job demands very differently. They assert that the literacy skills require more than the simple level of reading and writing that Dr. Goldstein contended is sufficient. Contrary to Dr. Goldstein's testimony, the Defendants presented testimony from job incumbents and supervisors which was thoroughly persuasive about the challenging reading and writing demands on DSP

Troopers. (*See, e.g.,* Tr. Vol. 3, 691:22– 694:20; 745:6-16; 798:7 – 800:13.) When asked to describe the difficulty of the reading material presented to Troopers, one of them answered, "it's difficult, because ... you're talking about the law, you're talking about abstract concepts that a person needs to read and analyze. It's not light reading that you can just grasp right away. You have to read it, think about it a little bit, go back, reread it." (Tr. Vol. 3, 692:4-10.) Perhaps it is because I too find the law challenging that I credit the assessment of those witnesses. Dr. Goldstein's observations about the literacy requirements of the Trooper job appeared to substantially overstate the simplicity of the demands and understate the importance of the skills.

92. However, I agree with Dr. Goldstein that the proposed 60% cutoff score would have resulted in a pass rate so high as to render the administration of the Alert basically useless. (*See* Tr. Vol. 5, 1472:8-23; Ex. 24 at p. 26; Ex. 25 at p. 7.) Yet we reach different conclusions from that fact. Dr. Goldstein tried to undercut the Defendants' use of the Alert by arguing that, because all Trooper candidates must have earned 60 college credits to be eligible for hire, they should be able to satisfactorily perform the reading and writing demands of the job. (Tr. 1301:6-13; Tr. Ex. 24, p. 29.) Dr. Goldstein conceded that he gathered no data and performed no analysis to support that conclusion. (Tr. Vol. 5, 1473:4-10.) He also acknowledged that he did not take into account the schools at which applicants earned their credits (Tr. Vol. 5, 1473:11-18), or the courses they took to obtain their credits, or the grades they received. Given the demands of the job and the potential variability of skills even among those with the requisite college credits, I find that it was sensible for the Defendants to seek some assessment tool apart from the requirement of 60 hours of college credit. Having

chosen the Alert as that tool, it was appropriate for the Defendants to set a cutoff score that did not make passing the test meaningless as a screening device, so long as that score was consistent with measuring the minimal literacy skills necessary for the job.

93. In determining the minimal literacy skill level necessary for the job, it was also appropriate for the Defendants to bear in mind the public safety consequences of setting a cutoff score too low. The Chief Deputy Attorney General of the State, a prosecutor with more than twenty years of experience in law enforcement, testified that police reports that are unclear, inconsistent, or incomplete can compromise or destroy the State's ability to prosecute criminals. (*See* Tr. Vol. 3, 818:3 – 821:10.) Bearing out what another witness had said (*see supra* at ¶ 17), he stated, "[i]t's almost as if if it's not in the police report, it did not happen." (Tr. Vol. 3, 819:24 – 820:2.) The fully legitimate concern that the public safety function performed by the DSP not be undermined by inadequate literacy skills is a proper factor to consider in determining what constitutes the minimum skill level for reading and writing and what Alert score properly reflects that minimum skill level.

III. CONCLUSIONS OF LAW

1. To the extent that any of my findings of fact may be considered conclusions of law, such findings are incorporated herein.

2. Jurisdiction over this case is proper under 28 U.S.C. § 1345 and 42 U.S.C. § 2000e-6(b). Venue is proper in this district under 28 U.S.C. § 1391(b).

3. Because of my previous conclusion that the Defendants' use of the Alert had an adverse impact on African American candidates for the job of DSP Trooper (*see* D.I. 261), the burden at trial was upon the Defendants to prove that their use of the Alert

was "job related for the position in question and consistent with business necessity." 42 U.S.C. § 2000e-2(k)(1)(A)(i).

4. Because the employment practice at issue is the use of a discriminatory cutoff score on an entry-level employment test, the Defendants could meet their burden of proof only by showing that the cutoff scores they chose measured "the minimum qualifications necessary for successful performance on the job in question." *Lanning I*, 181 F.3d at 481 (reiterated in *Lanning II*, 302 F.3d at 287).

5. The Uniform Guidelines on Employee Selection Procedures (the "Guidelines"), which are set forth in the Code of Federal Regulations state that,

> [r]eliance upon a selection procedure which is significantly related to a criterion measure, but which is based upon a study involving a large number of subjects and has a low correlation coefficient will be subject to close review if it has a large adverse impact. Sole reliance upon a single selection instrument which is related to only one of many job duties or aspects of job performance will also be subject to close review.

29 C.F.R. § 1607.14. Thus, while their burden of proof in this civil action is proof by a preponderance of the evidence, the evidence the Defendants have relied upon is properly subject to the close scrutiny called for by the Guidelines, both because the Alert was demonstrated to have generally low correlations to the requisite literacy skills for the Trooper job, as measured by the PDRF Composite, (*supra* at ¶ 42) while having a relatively large adverse impact (*supra* at ¶ 49) and because the Alert was used as a pass-fail hurdle, which made it, in effect, a "single selection instrument" upon which the Defendants solely relied.[55]

---

[55]The Defendants protest that they did not rely solely upon the Alert to hire Troopers, that instead the Alert was part of a "comprehensive selection process that assessed numerous skills, abilities and personal characteristics relevant to Trooper

6. I am also mindful that the Third Circuit has provided an explicit warning about statistical studies such as those performed for the Defendants in this case: "studies done in anticipation of litigation to validate discriminatory employment tests that have already been given must be examined with great care due to the danger of lack of objectivity." *Lanning I*, 181 F.3d at 481 (citing *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 433 n. 32 (1975) ("Studies so closely controlled by an interested party in litigation must be examined with great care.")).

7. I am required by Title VII and Third Circuit precedent to apply a two-pronged test to the Defendants' use of the Alert. Since use of the Alert has been shown to have a disparate impact on African Americans, the Defendants are required by the terms of Title VII "to demonstrate that the challenged practice is job related for the position in question and consistent with business necessity[.]" 42 U.S.C.A. § 2000e-2(k)(1)(A). The Third Circuit's opinion in *Lanning I* emphasizes that "Congress chose the terms 'job related for the position in question' *and* 'consistent with business necessity[,]'" 181 F.3d at 489 (emphasis in original), and that undue focus on one of those two prongs would impermissibly write the other out of the statutory test. *Id.* While the two prongs are distinct, they obviously bear a close connection to one another.

---

training and job performance." (D.I. 303 at p. 14.) The reality was, however, that candidates never advanced to the rest of the comprehensive selection process if they failed the Alert. In the only meaningful sense, that pass/fail hurdle was the sole selection instrument for the candidates who failed.

8. As to the requirement that the use of the Alert be "job related for the position in question," I conclude that it is.[56] The Plaintiff has acknowledged (D.I. 302 at 44), and I have found (*supra* at ¶ 7), that the skills measured by the Alert are related to the DSP Trooper job. There is a statistically significant correlation between performance on the Alert and performance in the literacy sphere of a Trooper's responsibilities. (*Supra* at ¶¶ 41, 42.) In and of itself, the use of the Alert was not problematic, but the manner of its use was. In that regard, the degree of the Alert's job-relatedness is relevant to the question posed by the second prong of the Title VII test for liability. In other words, the degree of validity of the Alert and the strength of its predictive power are relevant to determining whether the Defendants have met their burden of demonstrating that the way they used the test is consistent with business necessity.

9. Given my findings that the Alert has some unmeasured degree of content validity (*supra* at ¶ 27) and generally low criterion validity (*supra* at ¶ 42), it follows that the test must be used with particular care, since its predictive power is not great. *Cf. Ensley Branch v. Seibels*, 616 F.2d 812, 818 n. 16 (5th Cir. 1980) (affirming district court finding that test with low correlation coefficient "showed that the police test is predictive

---

[56]To say that the Alert itself is job-related may not answer the more precise question required by the statute, i.e., whether the practice of using the Alert as an entry level screening test with a cutoff score of approximately 75% is job-related for the position of Trooper. Even posed in that fashion, however, I conclude that the practice of administering the Alert is job-related, in the same sense one might say that a piece of evidence reaches the threshold of being relevant, without concluding anything further about the degree of relevance or the practical impact of the evidence. The United States concedes that the "job related" prong of the test has been satisfied in this case. (*See* D.I. 302 at 44.)

of better job performance, but that the magnitude of the positive prediction is so low that the test is worthless for all practical purposes."). That conclusion bears on the business necessity of using the Alert.

10. The Third Circuit in *Lanning II* further explained the standard for testing business necessity. Having said in *Lanning I* that the "business necessity" requirement is met if a "discriminatory cutoff score measures the minimum qualifications necessary for successful performance of the job in question[,]" 181 F.3d at 489, the Court in *Lanning II* accepted the lower court's implicit holding that the language "minimum qualifications necessary" means "likely to be able to do the job," 308 F.3d at 291.

11. That does not mean that the cutoff score on a discriminatory test, such as the Alert, should be set so that the predicted rate of job success for individuals who pass is 100%. Indeed, the Court stated that such a conclusion "would clearly be unreasonable[.]" *Lanning II*, 308 F.3d at 292. Nor do I take "likely" in this context to mean simply "more likely than not." The public certainly has a right to expect better than a 51% chance that members of its police force will be able to read and write well enough to do their job. Ample respect must be paid to the public safety concerns implicated by the literacy demands of the Troopers' job. As the Court stated in *Lanning II*, "police officers and the public they serve should not be required to engage in high-stakes gambling when it comes to public safety and law enforcement." *Id.* Because the stakes are high, I think it fair to say that "likely to be able to do the job" must be understood as meaning a high likelihood of being able to do the job.

12. When the evidence is viewed in that way, and with the close scrutiny warranted under the circumstances, it is clear that the Defendants have failed to carry

their burden of proof. The evidence presented by the Defendants fails to demonstrate that an applicant who scores below a standardized 75% on the Alert is unlikely to be able to perform the DSP Trooper job. To the contrary, convergent evidence shows that a very large number of applicants who score below 75% on the Alert are highly likely to be able to do the job.

13. Without reiterating all of my findings in this case, I note the following as particularly persuasive to me in reaching that conclusion. First is Dr. Wollack's testimony about the Alert. In the past, Dr. Wollack had recommended an Alert cutoff score as low as 62.5% (*see supra* at ¶ 49), yet no one presented evidence in this case or even suggested that police officers hired when that cutoff was in place have been sub-standard performers. Dr. Wollack also does not disagree with other jurisdictions' use of Alert cutoff scores that are lower than those used to screen candidates for DSP Trooper, even while acknowledging that the job demands on entry-level law enforcement officers are essentially the same throughout the country. (*See supra* at ¶ 7.) Furthermore, when looking specifically at the cutoff scores that were most recently used here in Delaware, Dr. Wollack conceded that, because of the standard error of measurement applicable to the Alert, applicants who failed the Alert by as many as 6.5 points may read and write just as well as applicants who passed. (*See supra* at ¶ 60.)

14. I have previously noted that the PDRF rating form and performance evaluation methodology used by Dr. Jeanneret as the basis for much of his statistical analysis has in it a fundamental assumption for which the Defendants failed to provide proof, namely that a rating below "Expected" on the reading and writing dimensions of Dr. Jeanneret's PDRF form is synonymous with, and was understood by the SMEs to

mean, “lacks the minimum reading and writing skills necessary to do the job.” (*See supra* at ¶ 37.) A great deal of the Defendants’ proof is entirely dependant on the assertion that an individual who received a rating below 144.5 on the PDRF Composite lacks the minimum reading and writing skills necessary to do the job, but the Defendants had to admit that rating supervisors were never asked about minimally acceptable performance. (*Id.*)

15. Even accepting the Defendants’ assumption in that regard, however, the use of reverse regression with truncated data is not worthy of credence because it is revealing only of what was foreordained: it is mathematically guaranteed to identify a cutoff score above the actual cutoff score used by the DSP. After the Defendants’ efforts to correct for the truncation problem so that the regression line dictated an Alert cutoff score in the range of 72-75%, one still is left to account for the Defendants’ failure to address the conditional distribution around the regression line. A cutoff score pegged at the point that correlates with a PDRF Composite rating of 144.5, which is the point the Defendants claim represents minimally acceptable performance, would eliminate 50% of the individuals who would be predicted to perform the job at that level of competence. That fact is brought to life by the 14 out of 18 Alert failers in the validation sample who went on to successfully perform the DSP Trooper job. Although the Defendants try to minimize that evidence by saying the failers who later passed may have improved their skills between testings, they didn’t present any proof of that, and it seems to me, in light of all the evidence, that the more likely reason for the different outcome on their later testings is the bluntness of the test instrument itself. While the

Alert has been shown to be discriminatory in the legal sense, it has relatively weak discriminatory power in a psychometric sense.

16. Based on the evidence presented, I hold that an Alert cutoff score of 75% does not correspond to the minimum qualifications necessary to perform the DSP Trooper job. Therefore, under the controlling law in this Circuit, the Defendants have failed to bear their burden of proving that their use of the Alert was "job related for the position in question and consistent with business necessity." *See* 42 U.S.C. § 2000e-2(k)(1)(A)(i); *Lanning I*, 181 F.3d at 481; *Lanning II*, 308 F.3d at 287.

17. Both parties agree that they would like me to provide some guidance as to what I view as the appropriate cutoff score or range of cutoff scores that reflect the minimum level of literacy required to perform the job of a DSP Trooper. (Tr. Vol. 6, 1597:21-1603:6; 1627:16-1628:19.) To that end, and in light of all the evidence, I believe that the range of cutoff scores on the Alert that one can reasonably argue corresponds to the minimum qualifications necessary to perform the DSP Trooper job is from 66% to 70%.

18. The purported objectiveness of the statistical evidence in this case seemed to melt away as well-respected, highly qualified statistical experts drew widely varying conclusions from the data. The Defendants' experts took the Plaintiffs' numbers and generated the highest possible cutoff score, while the Plaintiff's experts used the Defendants' numbers to achieve the lowest possible cutoff score. At the end of the day, however, the range of appropriate cutoff scores, as defined by the *Lanning* standard, appeared to me to form a band in the 66-70% range.

19. Identifying a range of cutoff scores for the Alert, of course, does not constitute a finding that all candidates scoring in or above that range were entitled to be hired as Troopers. The Alert, as the Defendants chose to use it, was a first screen, but candidates passing it may have been unsuitable for employment for any number of lawful reasons which the Defendants may have acted upon, had the Alert score not been excessive.

20. Finally, I reiterate something that I mentioned in my first Opinion in this matter (*see* D.I. 261 at 14-15), that the United States has readily conceded (*see* Tr. Vol. 6, 1622:8-21), and that is worthy of particular emphasis now that all of the evidence has been aired, namely that there is nothing to indicate that racial animus or an intent to discriminate motivated the Defendants in using the Alert. On the contrary, the evidence indicates that, in setting a cutoff score, the Defendants were following the advice of the author and vendor of the test and that they did so in an effort to be faithful to their obligations to be fair to applicants and to be guardians of the public interest in effective law enforcement. The good faith of the decision makers at the time, however, is not the issue in a disparate impact case such as this. Under Title VII and controlling precedent, the way the Defendants used the Alert was unlawful, regardless of motivation.