295 F.Supp.2d 1298 (2003)

Johnny REYNOLDS, et al., Plaintiffs,

v.

ALABAMA DEPARTMENT OF TRANSPORTATION, et al., Defendants.

No. CIV.A. 85-T-665-N.

United States District Court, M.D. Alabama, Northern Division.

December 19, 2003.

1299 *1299 Winn S. L. Faulk, pro se, Daphne, AL, for special master.

Leonard Gilbert Kendrick, Richard H. Gill, Copeland, Franco, Screws & Gill, P.A., Florence Belser, pro se, Montgomery, AL, Robert F. Childs, Jr., Richard J. Ebbinghouse, Ann K. Wiggins, Abigail P. van Alstyne, Robert L. Wiggins, Jr., Jon C. Goldfarb, Gregory O. Wiggins, Russell W. Adams, Deborah A. Mattison, Rocco Calamusa, Jr., Rebecca Anthony, Kimberly C. Page, Kell A. Simon, H. Wallace Blizzard, III, Eden J. Brown Gaines, Steven Atha, Susan Donahue, Wiggins Childs Quinn & Pantanzis, PC, Stanley W. Logan, Baker Donelson Bearman Caldwell & Berkowitz PC, C. Paige Goldman, David P. Whiteside, Jr., Law Offices of David P. Whiteside, Jr., Raymond P. Fitzpatrick, Jr., R. Scott Clark, J. Michael Cooper, Gary Lamar Brown, Fitzpatrick, Cooper & Clark, Birmingham, AL, Julian L. McPhillips, Jr., McPhillips, Shinbaum & Gill, Rick Harris, The Glassroth Law Firm, PC, Montgomery, AL, Claudia H. Pearson, Vestavia Hills, AL, William R. Murray, Northport, AL, for plaintiffs/ intervenors-plaintiffs.

R. Taylor Abbot, Jr., Spain & Gillon, L.L.C., Thomas R. Elliott, Jr., Allen R. Trippeer, Jr., C. Dennis Hughes, London & Yancey, William F. Gardner, William K. Thomas, Cabaniss, Johnston, Gardner, Dumas & O'Neal, Laura Ellison Proctor, Jacquelyn D. Smith, Erica L. Sheffield, Baker Donelson Bearman Caldwell & Berkowitz PC, Andrew P. Campbell, Jonathan H. Waller, David M. Loper, Amy L. Stuedeman, Cinda R. York, Campbell Waller & Poer LLC, Joseph L. Cowan, II, Nettles Hart Hess & Hughes, Lisa Wright Borden, Baker Donelson Bearman Caldwell & Berkowitz PC, Robert R. Baugh, David W. Long, David W. Long-Daniels, Sirote & Permutt, P.C., James M. Keel, London & Yancey, Anne R. Yuengert, Bradley, Arant, Rose & White, LLP, Roger L. Bates, Mark T. Waggoner, Hand Arendall, L.L.C., Jason Michael Osborn, Baker Donelson Bearman Caldwell, & Berkowitz PC, Chris Mitchell, Constangy, Brooks & Smith, Thomas O. Sinclair, Wendy T. Tunstill, Campbell Waller & Poer LLC, William P. Gray, Jr., Gray Johnston & Associates, Attorney at Law, Birmingham, AL, Stephen L. Scott, The Kullman Firm, New Orleans, LA, Robert M. Weinberg, William H. Pryor, Jr., Attorney General, Office of the Attorney General, Algert S. Agricola, Jr., Slaten & O'Connor, PC, David B. Byrne, Jr., Henry Clay Barnett, Jr., Christopher W. Weller, Mai Lan F. Isler, Capell Howard PC, Robert A. Huffaker, Rushton, Stakely, Johnston & Garrett, P.A., David R. Boyd, Balch & Bingham, Jim R. Ippolito, Jr., Jack Franklin Norton, Kenneth Lamar
1300 Thomas, *1300 Christina H. Jackson, Thomas Means Gillis & Seay PC, Alice Ann Byrne, Ronald Wayne Wise, Law Office of Ronald W. Wise, Troy R. King, Montgomery, AL, Patrick H. Sims, Cabaniss Johnston Gardner Dumas & O'Neal, Mobile, AL, Eric D. Hoaglund, McCallum Law Firm LLC, Vestavia Hills, AL, Champ Lyons, Jr., Point Clear, AL, for defendants.

Elaine R. Jones, Norman J. Chachkin, New York, NY, Barbara R. Arnwine, Thomas J. Henderson, Richard T. Seymour, Teresa A. Ferrante, Lawyers' Committee for Civil Rights Under Law, Washington, DC, for amicus.

OPINION

MYRON H. THOMPSON, District Judge.

In this longstanding lawsuit, African-American plaintiffs charged defendants Alabama Department of Transportation, Alabama State Personnel Department, and their officials, with racial discrimination in employment. This lawsuit is once again before the court, this time on the plaintiffs' motions for civil contempt relief

relating to the following five exams administered by the defendants: (1) Civil Engineer-Construction Option Examination; (2) Senior Right-of-Way Examination; (3) Civil Engineer Manager Examination; (4) Civil Engineer Administrator Examination; and (5) Civil Engineer-Design Option Examination. For the reasons stated below, the court will deny the plaintiffs' motions.

I.

This lawsuit, filed in 1985, charged the defendants with widespread and long-lasting racial discrimination in the Alabama Transportation Department. The plaintiffs based this lawsuit on the following: Title VII of the Civil Rights Act of 1964, as amended, 42 U.S.C.A. §§ 1981a, 2000e through 2000e-17; the fourteenth amendment to the United States Constitution, as enforced by 42 U.S.C.A. § 1983; and 42 U.S.C.A. § 1981. The jurisdiction of the court has been invoked pursuant to 28 U.S.C.A. § 1343 (civil rights) and 42 U.S.C.A. § 2000e-5(f)(3) (Title VII).

In 1994, the parties to this case entered into a consent decree providing for extensive and complex remedial relief. *Reynolds v. Alabama Dep't of Transp.,* 1994 WL 899259 (M.D.Ala.1994). Included in that consent decree are the provisions now at issue, which concern the instruments the defendants may use to select candidates to fill job openings. Specifically, the plaintiffs allege that the defendants have not complied with ¶¶ 4(a) and 8 of Article Three, which state in pertinent part,

> "4. Validation of criteria:
>
> (a) Personnel will develop and thereafter use only selection criteria and procedures that have been validated in accordance with the Uniform Guidelines on Employee Selection Procedures.
>
> . . .
>
> "8. During the selection of examination type and development of the selection instrument, [the State Personnel Department] will search for effective alternative devices which would have lesser disparate impact. Where a selection device shows substantial disparate impact upon use, [the Personnel Department] will search for effective alternative devices which would have lesser disparate impact in future selection decisions, and will utilize such devices unless impracticable; ...."

Id. at *8. The plaintiffs bear the burden of proving by *clear and convincing evidence* that the defendants are in violation of the consent decree. *Reynolds v. McInnes,* 338 F.3d 1201, 1211 (11th Cir.2003).

1301 The plaintiffs make two analytically distinct arguments for why these five tests, as *1301 developed, do not comply with these consent decree provisions. First, the plaintiffs argue that the exam scores are not weighted in a way that minimizes adverse impact while maintaining (or increasing) the content validity of the exams; second, they argue that the tests are not sufficiently valid to be used for the purpose of rank-ordering candidates' scores. In response, the defendants argue that the exams at issue are highly content valid, thus allowing for the rank-ordering of candidates' scores, and that the plaintiffs have not met their burden of proving the existence of an alternative score-weighting method that would both be as content valid as and have lower adverse impact than the defendants' method.[1]

II.

A. Content Validity

Under ¶ 4(a) of Article Three of the consent decree, the defendants may use only selection criteria that have been validated in accordance with the Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. §§ 1607.1 to 1607.18. The Uniform Guidelines provide for different methods by which a party can validate a selection device; the defendants have attempted to validate the exams at issue using content validation. Content validation is a process by which the user of a selection procedure "should show that the behavior(s)

demonstrated [in that procedure] are a representative sample of the behavior(s) of the job in question or that the selection procedure provides a representative sample of the work product of the job." 29 C.F.R. § 1607.14C(4). In other words, as the parties' experts explained, content validation looks at the process by which one constructs a selection procedure so that the content of that selection procedure corresponds to the content of the job; as applied here, content validation focuses on the procedures the defendants used to map the content of each of the five jobs at issue into the content of the exams used to select candidates for those jobs.

Notably, content validity is not an all or nothing proposition; rather, there is a continuum of levels of content validity. Where a selection procedure falls along that continuum determines the purposes for which it can be used; for example, "[e]vidence which may be sufficient to support the use of a selection procedure on a pass/fail (screening) basis may be insufficient to support the use of the same procedure on a ranking basis." 29 C.F.R. § 1607.5G. Additionally, as the plaintiffs' experts conceded, more adverse impact is tolerable in selection procedures that are highly content valid.

B. Tests at Issue

1. Process of test development

The defendants followed essentially the same process in developing each of the five tests at issue. For each job classification, the State Personnel Department first performed a job analysis. The purpose of this analysis is to describe the job and what happens on the job, focusing on those activities that are important and are done with some frequency. The analysis also breaks down the job into its component knowledges, skills, and abilities, often called KSAs. As an example, a KSA might be "knowledge of trigonometry," "skill in driving," or "ability to research information."

Those KSAs identified in the job analysis were then narrowed to form a job-content domain, which is, essentially,
1302 a *1302 profile of what is done on the job. The Personnel Department developed the job-content domain by using subject-matter experts (individuals who have actual job experience in the job being studied and who are knowledgeable of its various requirements), called SMEs, to evaluate the KSAs identified in the job analysis. Whenever possible, the Personnel Department gave special attention to selecting minorities and females for participation as SMEs. The SMEs answered a number of questions about each KSA, questions designed to determine both the relative importance of the KSAs and whether a KSA was necessary at entry.

Specifically, each SME was first asked whether he performed a given KSA. If the SME answered no, he was instructed to move on to the next KSA. If the SME answered yes, he was then asked to indicate the importance of the KSA on a five-point scale, from not important (0) to crucial (4), and to make a yes or no judgment as to whether the KSA was necessary upon entry to the job. The job-content domain was made up of those KSAs that were performed by at least 67% of the SMEs and that had a mean importance rating of at least 2.0.

The next step in the process further narrowed the KSAs eligible for testing by eliminating all of those that were not rated as necessary-at-entry by at least half the SMEs who said they performed that KSA. As noted above, however, if an SME indicated that he did not use a given KSA, he was not asked to provide a response to the question of whether that particular KSA was necessary-at-entry. Thus, the 50% necessary-at-entry screen means that a KSA was eligible for testing if half of the SMEs who said they performed that KSA also rated it as necessary-at-entry. That a KSA was found to be eligible for testing does not mean that half of all SMEs rated it as necessary-at-entry. The group of KSAs that survived through this step in the process was the "job-content domain eligible for testing."

Those KSAs that were found to be eligible for testing were then whittled down by the test developers to those actually used in the exam. There were a number of reasons that a KSA was not be used on an exam. For example, in exams that were open to external candidates, any KSA that tested Transportation Department-specific material was excluded. Other KSAs were excluded because of feasibility issues they were not so important that the exam should have an added component to measure them, incurring extra time and expense for both the Transportation Department and the candidate taking the test. In some exams, KSAs were excluded

because the possession of that KSA was assessed through the measurement of a different KSA; for example, the "ability to learn trigonometry" KSA was not assessed in an exam that already measured "knowledge of trigonometry." The KSAs that survived this process, and thus were measured in the exam, were the test-content domain.

After the test-content domain was established, the test developers needed to decide what type of exercise to use to evaluate those KSAs. For all of the exams at issue, the test developers chose to use work-sample exams. Work-sample exams simulate work that a person would actually perform on the job using job-related scenarios; they "are designed specifically to mirror a given job in question."[2] It is generally agreed by the parties' experts that work-sample exams are expected to have the highest level of content validity of any exam type.

1303 *1303 The specific components of each of the work-sample exams at issue were developed by the SMEs with guidance from the test developers. The SMEs identified critical incidents that typically occurred in the job that required possession of the KSAs to be measured in the exam, and then they developed exercises to assess those KSAs. Throughout this process, the test-developers emphasized to the SMEs that the exercises should focus on those situations that employees would be expected to handle upon entry to the job and not on those situations which required experience to handle effectively.

In the exam for Civil Engineer-Construction Option, for example, this process resulted in four work-sample exercises: (1) a construction-plan-reading exercise, in which the candidate would review roadway plans, survey notes, and other technical documents and use that information to respond to questions, identify errors, and make computations; (2) a role-play exercise, in which the candidate played the role of a civil-engineer-project engineer dealing with a contractor; (3) a payment-voucher review and memo-writing exercise, in which the candidate was given a number of items related to a payment voucher and was required to review that information, determine if there were any errors in the invoices, and respond in writing; and (4) a scheduling/in-basket exercise, in which the candidate was required to complete a work schedule for the coming week, assigning employees to specific projects and tasks.

After creating the exercises, the test developers then worked with the SMEs to develop objective criteria by which to rate candidates' performance on those exercises. At this point, the exercises were put into final form and were sent to the parties' industrial and organizational psychologists for review. The exams then went to a group of SMEs for a final review and revision. In that review, the SMES were asked to answer the following questions (with answers in parenthesis): (1) Is this exercise job related? (1 = yes, 0 = no). (2) What is the quality of each exercise?[3] (0 = not job related, 1 = too easy, 2 = too difficult, 3 = ambiguous, 4 = inaccurate, 5 = biased, 6 = good item). (3) To what extent would the ability to respond to this exercise distinguish between superior and adequate levels of competence in the job being evaluated?[4] (0 = not at all, 1 = slightly, 2 = moderately, 3 = to a great extent). (4) Indicate to what extent each performance dimension[5] would be helpful in responding to each question. (0 = not assessed/measured/tapped, 1 = somewhat helpful, 2 = very helpful, 3 = not necessary)."[6] The purpose of these questions was to ensure that the content of the exams was appropriate. No exercise in the exams at issue was rated by an SME as biased, inaccurate, ambiguous or not job-related.

1304 *1304 ## 2. Test administration

For these exams, not all candidates who wanted to take the test were allowed to do so. Instead, candidates had to demonstrate possession of certain minimum qualifications, called MQs, in order to be able to take each of the tests at issue.[7]

Approximately three weeks before the administration of each of the exams at issue, those candidates who had passed the relevant MQs were sent a "how to prepare" manual to assist in their preparation. This manual informed the candidates about the exam process; told them what to expect, what to bring, and what KSAs were measured; described each exercise format; told them how the exercises would be scored; and gave the candidates tips on how to prepare for the exam. In part, the purpose of this manual was to reduce the exams' adverse impact on minority candidates.

The exams at issue have all been administered and scored at least once. Because of the subjective nature of a candidates's performance on most work-sample exercises, any aspect of an exam that required a grader to use his judgment in deciding the candidate's score was graded by a panel of two graders. Where possible, the test-developers sought panels with one black and one white grader, one female and one male grader. In some cases, this required bringing in qualified graders from out of state. These measures, too, were taken in order to minimize the exams' adverse impact on minority candidates.

While the plaintiffs have some complaints about the process of test development, the plaintiffs' experts approved all of the exams at issue through this stage in the process; specifically, the plaintiffs' experts approved of the content of the exams and the choice of work-sample exams. The plaintiffs' experts take issue with the way the exams scores were ultimately weighted, but they found the content of the exams themselves to have a good resemblance to the content of the jobs at issue. For example, Dr. Joel Lefkowitz, one of the plaintiffs' experts in industrial and organizational psychology, testified that the exam exercises "are designed in a general way to represent meaningful aspects of the job. We don't dispute that. You look at the examinations, they bear quite a decent resemblance to the jobs they're meant to be selecting people for."[8] Similarly, Dr. James Outtz, another of plaintiffs' experts, testified that he approved both the job analyses and the content of the tests.[9]

3. Test scoring

After the five exams at issue were administered and graded by the panels, the test-developers had raw scores from each exercise within the exams, which they gave to the State Personnel Department. These exercises were graded on different scales, which had to be standardized before they could be added together; some exercises had multiple components that yielded multiple scores. The plaintiffs take no issue with the calculation of the candidates' raw scores for each exercise or the standardization of those scores.[10]

It is the next step in the process that is the focus of the parties' controversy. After standardizing the scores from each exercise, the defendants then summed those scores using unit weighting. In other words, after
1305 standardization, the defendants multiplied the score from each exercise by one, and then added those scores
*1305 together, resulting in a single total score for each candidate.[11] A linear transformation was then used so that each candidate's score would fall between 70 and 100, and so that the candidates' scores could be rank ordered.

While the parties dispute the correct term—"potential" or "actual" adverse impact—for the differences between black candidates' scores (as a whole) and white candidates' scores (as a whole) on the exams at issue, it is undisputed that the scores of black and white candidates were different on these exams. Not only did black and white candidates' composite scores differ, but their scores also differed on each exercise within the exams, sometimes at a statistically significant level. Generally, blacks candidates' scores were lower than those of white candidates—although in some instances the black candidates did better on a particular exercise or aspect thereof—and black candidates' composite scores were lower than white candidates' composite scores on all five exams at issue.

4. Ultimate selection

Because of the plaintiffs' objections, the ultimate selection of employees from these exams has not yet taken place. When it does occur, however, the process will proceed as follows. The candidates' scores will be placed on a continuous register, meaning that they will be added to a list along with the scores of candidates who have previously taken the exam, a register to which names are added even when there is no opening in the job at issue.

When a position opens up, a certificate of eligibles, called a COE, is issued from the appropriate register. A candidate's test score is only one factor in determining whether his name appears on the COE for the available position. Other factors include: whether the candidate is willing to work in the geographic region in which the job

is located; whether the candidate is willing to do shift work, if required for the available position; whether the candidate is willing to travel overnight, if required; whether the candidate is willing to work part-time/full-time, as needed; and the veterans' preference.[12] The COE that issues is comprised of those candidates with the top ten scores on an exam, plus ties, taking into consideration the factors listed above. Once the COE issues, the Transportation Department may interview those people whose names appear thereon, and then hires someone from the COE. Save for those circumstances in which the *Frazer/Ballard* no-bypass rule applies, the Transportation Department may hire anybody on the COE, even a candidate whose test score is lower than that of others who appear on the COE.[13]

1306 *1306 **III.**

A. Necessary-at-Entry Dispute

The plaintiffs' first objection to the defendants' testing procedures revolves around the defendants' calculation of the necessary-at-entry screen for KSAs. The plaintiffs argue that the defendants erred by calculating the 50% necessary-at-entry screen with no regard for those SMEs who said they did not use a KSA. Rather than using KSAs that were rated necessary-at-entry by half of the SMEs who said they performed that KSA, according to the plaintiffs, the defendants should have used only KSAs that were rated as necessary-at-entry by half of all SMEs.

Insofar as the plaintiffs would have the court find that this method of calculating the necessary-at-entry screen is in violation of the parties' content decree or the Uniform Guidelines, the court refuses to so find. In a report and recommendation adopted by the court in full, Special Master González addressed and rejected these same arguments by the plaintiffs.[14] This matter has already been decided against the plaintiffs, and the court declines to reconsider it here.

The plaintiffs have also suggested, however, that the defendants' calculation of the necessary-at-entry screen dispute is a reason for why the court should find the defendants in contempt of the consent decree; specifically, the plaintiffs argue that the defendants' calculation of the necessary-at-entry screen lowers the content validity of the exams at issue. Insofar as the plaintiffs' necessary-at-entry arguments go to the content validity of the exams, the court will address those arguments below.

B. Content Validity of Exams Using Unit Weighting

Central to the parties' dispute is the content validity of the exams at issue using the defendants' score-weighting method, unit weighting. The plaintiffs argue that the exams are not content valid enough to be used for the purpose of rank-ordering candidates' scores, and that unit-weighting method does not minimize adverse impact while maintaining (or increasing) the content validity of the exams, both of which violate the consent decree. The defendants disagree, arguing that the exams are highly content valid as they are now scored. Taking the plaintiffs' arguments one step at a time, the court will first address the content validity of the exams using unit weighting.

As discussed above, content validation focuses on the procedures the defendants used to map the content of each of the five jobs at issue into the content of the exams used to select candidates for those jobs. Content validity is not an all or nothing proposition; rather, as stated, it is a continuum. The court must determine where on that continuum these exams fall. In so doing, the court relies heavily on the parties' experts' testimony, as there are few, if any, objective measures by which one might gauge the content validity of an exam.[15] Instead, content validity appears to be a very subjective measure, a sort of "you know it when you see it" comparison of the content of an exam with the content of the job and analysis of the procedures used to get from one to the other.

1307 *1307 After having heard three days of expert testimony, read expert reports and expert depositions, and looked at the defendants' content-validation reports, the court concludes that the exams at issue are highly content valid. First, all of the experts agreed that, in general, work-sample exams (which were used here) are highly content valid. As Dr. Lefkowitz explained, work-sample exams "are designed specifically to mirror a given job in question,"

[16] allowing for the testing of skill dimensions that occur naturally together.[17] Indeed, the Uniform Guidelines explain that "[t]he closer the content and the context of the selection procedure are to work samples or work behaviors, the stronger is the basis for showing content validity." 29 C.F.R. § 1607.14C(4). The defendants' expert, Dr. Philip Roth, even testified that there is literature on work-sample exams that suggests those exams have "the highest level of predictive validity."[18]

Second, the plaintiffs' experts approved these five exams specifically. While the parties dispute the extent of that approval, it is undisputed that the plaintiffs' experts approved of the relationship between the content of these exams and the content of the underlying jobs. Dr. Lefkowitz testified that these are good exams "with regard to the examination content and the extent to which the exam content mirrors ... the job."[19] The way in which these exams were developed supports this conclusion. The defendants extensively used SMEs familiar with the underlying jobs, both to choose the KSAs to be measured in the exam and to develop the exercises to assess those KSAs. After those exercises were made final, no SME rated any exercise as biased, inaccurate, ambiguous, or not job-related.

Third, the court finds that the defendants' method for calculating the necessary-at-entry screen does not negatively affect, to any noticeable degree, the content validity of the exams at issue. The defendants' experts testified that their approach was one of two professionally acceptable methods; according to the defendants, one either asks just those SMEs who use a KSA whether that KSA is necessary-at-entry, or one asks all SMEs whether a KSA is necessary-at-entry.[20] No party presented evidence at the hearing before the court as to the comparative efficacy of these two approaches; presumably, the defendants had some reason for rejecting the approach that would allow the test developers to find out what each SME thought about whether a KSA is necessary-at-entry, and adopting one whereby they would find out only whether the SMEs who use a KSA thought it was necessary at entry.

Instead of addressing these two methods of calculating the necessary-at-entry screen, the plaintiffs argued that the defendants should have used a third method. Under the plaintiffs' approach, one assumes that an SME who says that she does not use a KSA would also say that KSA is not necessary-at-entry.[21] The defendants' experts
1308 disagree with this method of calculating the necessary-at-entry screen, arguing that this assumption is *1308
invalid.[22] The court agrees. As Special Master González concluded, the assumption that an SME who says that he does not use a KSA would also say that KSA is not necessary-at-entry "is unsupported by the evidence. There is no evidence that a SME who said he or she did not use a KSA at entry was intending to say that as a consequence a particular KSA is not needed at entry. It is quite possible that persons who have been in a job for some years reach the point where KSAs that were once necessary are not now required."[23]

Insofar as the plaintiffs would have the court find the exams at issue less content valid because the defendants did not calculate the necessary-at-entry screen using plaintiffs' method, the court must reject the plaintiffs' argument. The plaintiffs' method for calculating the necessary-at-entry screen makes an assumption that is unsupported by logic or the evidence, and the defendants' cannot be faulted for refusing to adopt that method. While it is unclear whether or not the defendants chose the best method for calculating the necessary-at-entry screen, it is clear that they chose a professionally acceptable method, one that does not reduce the content validity of the exams at issue.

Fourth, and probably most importantly, the court finds that the defendants' choice to score these exams using unit weighting does not diminish their content validity. The plaintiffs argue that the defendants' choice of unit weighting severely reduces the content validity of these exams because, under unit weighting, there is no relationship between (a) the number of KSAs measured by an exercise, the extent to which those KSAs are measured in an exercise—for example, if a KSA was emphasized in an exercise rather than being used briefly—and the relative importance of those KSAs to the job[24] and (b) the weight of that exercise relative to the weight of other exercises within the exam. The plaintiffs argue that regardless of how many KSAs an exercise tests and how important those KSAs are, the defendants' choice of unit weighting gives that exercise the same weight as an exercise that tests more (fewer) KSAs that are more (less) important to the job. The plaintiffs say that the defendants chose unit weighting because it was convenient, not because it related to the validity of the exams in any way.

The defendants, on the other hand, make a number of arguments in defense of their choice of unit weighting. First, the defendants' expert report points out that "the nature or strategy for measurement of a work sample test is to assess dimensions as they tend to occur on the job. It is quite possible for multiple dimensions to naturally
1309 occur together."[25] If these work-sample exams were weighted on the basis of the number or importance of the *1309 KSAs in each exercise — essentially treating each KSA as an individual unit—it would ignore the way in which those KSAs interact on the job. Given this interaction of KSAs on work-sample exams, Dr. Maury Buster, another of the defendants' experts, testified that the plaintiffs' proposed weighting schemes would be more appropriate for written-type exams than for work-sample exams.[26]

Second, the defendants' experts point out that unit weighting allows for scores that represent a balance of both content and communications skills, and that it is a compensatory scoring system, in that a candidate who does poorly on one exercise can compensate by doing well on the other exercises.[27] By comparison, a system based on KSA counting or importance might result in exercises having significantly different weight, meaning that content and communications skills might not be reflected in the final score in a balanced fashion, and that a candidate who does poorly on a heavily weighted exercise may not be able to compensate for that, no matter how well he does on other exercises. Third, Dr. Philip Bobko, another of the defendants' experts, testified that the defendants chose unit weighting "to be true to the exercises and scores that we were given by the test developers."[28] By this, he meant that the test developers were trying to use many different methods for assessing the capabilities of individual applicants—such as using written materials, personal interaction, calculators and other mathematical material—and that unit weighting was chosen because it would preserve that multiplicity of different methods.[29]

Finally, the defendants argue that unit weighting is both acceptable and highly content valid because it produces scores that are highly correlated with those under systems that base their weights on the underlying KSAs. As both Drs. Bobko and Outtz testified, academic literature has shown that there is a high correlation between scoring systems that use different weighting methods,[30] correlation so high that, when candidates are ranked based on those scores, the ranks would change very little, if at all, under the different systems.[31] Furthermore, Dr. Buster testified that he performed comparisons of three different weighting systems—unit weighting and two others based on the KSAs underlying each exercise—on a number of Transportation Department tests, including the Senior Right-of-Way Specialist, Civil Engineer-Design Option and Civil Engineer-Construction Option exams at issue here, and found those weighting systems to be extremely highly correlated, on the magnitude of .99 and above on a 1.0 scale.[32]

In conclusion, taking all of this together —(1) work-sample exams are highly content valid in general; (2) the parties' experts agree that the content of these exams specifically are highly representative of the content of the underlying job; (3) the exams were developed using extensive input from SMEs familiar with the underlying jobs, who both chose the KSAs to be measured and developed the exercises to assess those KSAs; (4) the defendants' method of calculating the necessary-at-entry screen is a professionally accepted practice that does
1310 not detract from the *1310 content validity of the exams in any significant way; and (5) the significant correlation between unit weighting and other scoring systems, including those based on the underlying KSAs—the court finds that, as a result of the procedures used by the defendants, the content of these five exams is highly representative of the content of the underlying job, and the defendants use of unit weighting does not detract from that representativeness. In other words, the exams at issue are highly content valid.

C. Plaintiffs' Alternate Use Argument

Although the court has found the five exams at issue to be highly content valid, that finding does not resolve the plaintiffs' contempt motions. Not yet settled is the main thrust of the plaintiffs' arguments— that the defendants are in violation of the consent decree because they adopted a score-weighting method that does not minimize adverse impact while maintaining (or increasing) the content validity of the exams, in violation of the requirements of consent decree.

1. Burden of proof

As discussed in a previous appeal in this case, on a motion for a finding of contempt, the burden is on the moving party (here, the plaintiffs) to prove by clear and convincing evidence that the non-moving party (here, the defendants) is in violation of the consent decree. *Reynolds v. McInnes,* 338 F.3d 1201, 1211 (11th Cir.2003). The plaintiffs, however, argue that the consent decree in this case changes that burden for the motions at issue. Essentially, the plaintiffs would construe ¶ 8 of Article Three of the consent decree to require the defendants to prove that they have used the device with the least disparate impact possible.

The applicable language of ¶ 8 reads:

> "8. During the selection of examination type and development of the selection instrument, [the State Personnel Department] will search for effective alternative devices which would have lesser disparate impact. Where a selection device shows substantial disparate impact upon use, [the State Personnel Department] will search for effective alternative devices which would have lesser disparate impact in future selection decisions, and will utilize such devices unless impracticable; ...."

The plaintiffs' argument, that this paragraph puts the burden on the defendants to prove that they have used the device with the least disparate impact, ignores the provision's plain language. First, ¶ 8 addresses selection devices that show "substantial disparate impact *upon use.*" (Emphasis added). While the five exams at issue have been administered, there is no evidence in the record of their "use"; in other words, there is no evidence that certificates of eligibles have been issued or selections made based upon these exams. Any other interpretation of the word "use" in this context such as allowing "use" to mean simply the administration of the exams would unduly constrain the meaning of the word.

Second, and more important, no matter what the meaning of the word "use," the plaintiffs' argument must be rejected because it ignores the provision's requirement that, if a device shows substantial disparate impact, the defendants must "search for effective alternative devices which would have lesser disparate impact in *future* selection decisions." (Emphasis added). In other words, if the defendants employ these exams, and they show substantial disparate impact, ¶ 8 requires the defendants to search for alternative devices and to use those
1311 devices for later selections, unless impracticable.[33] This article *1311 of the consent decree in no way prevents the defendants from ever using these exams; it does not stop them from using the exams before any selections are made.

The plaintiffs also argue that their interpretation must be adopted to avoid rendering ¶ 8 meaningless. They say that the defendants are already required by ¶ 4 of Article Three to use only selection criteria that have been validated in accordance with the Uniform Guidelines, and that the Uniform Guidelines require the defendants to use a device with lesser adverse impact if that device is "substantially equally valid." 29 C.F.R. § 1607.3B. The plaintiffs argue that ¶ 8 removes any "substantially equally valid" requirement, and instead requires the defendants to use a device with lesser adverse impact "unless impracticable."

While this argument has some merit, it again ignores the fact that ¶ 8 requires the defendants to utilize devices with lesser adverse impact "in future selection decisions" only. As the exams at issue have not yet been used for any selection decisions, the plaintiffs' reading of ¶ 8 cannot stand. This provision of the consent decree clearly allows the defendants to use a valid-selection procedure to see if it shows substantial disparate impact; only if that device "shows substantial disparate impact upon use" must the defendants utilize alternative devices with less disparate impact. As such, the court cannot agree with the plaintiffs' reading and must reject their burden-shifting argument. Paragraph 8 does not require the defendants to prove that they used the selection device with the least disparate impact, and the burden remains on the plaintiffs to show by clear and convincing evidence that the defendants are in contempt of the consent decree.

2. Plaintiffs' alternative uses

The court now turns to the heart of the plaintiffs' argument: that the defendants are in contempt of the consent decree because they have not complied with the Uniform Guidelines. The plaintiffs argue that the defendants have not complied with the Uniform Guidelines requirement that, "Where two or more selection procedures are available which ... are substantially equally valid for a given purpose, the user should use the procedure which
1312 has been demonstrated to have the lesser adverse impact." 29 C.F.R. § 1607.3B.[34] In *1312 response, the defendants argue that the plaintiffs have not shown either that their proposed alternatives—a different weighting scheme and banded scoring—are substantially equally valid to, or that they have less adverse impact than, the tests as the defendants would use them.

a. KSA-based weighting

As a preliminary matter, the court notes that the plaintiffs have never explained exactly how they would have the defendants weight the exam scores. Instead, the plaintiffs' experts have stated very generally that they would weight scores in a way that considers both the KSAs underlying an exercise—presumably their number, relative importance, and the extent to which they are measured—and the d-statistic of each exercise.[35] The d-statistic of an exercise is a statistical measure that quantifies in a standardized way the differences in scores between two groups (here the black applicants (as a group) and the white applicants (as a group) who took an exam), thereby showing which group scored better, the size of that scoring differential, and whether that difference is statistically significant. For these exams, the defendants have calculated the d-statistic both for each exercise individually—and at the component level, if an exercise produces more than one score—and for each exam as a whole. While the parties debate the feasibility of such a system (in particular, how one may develop stable weights given that the d-statistic will change with each exam administration, and the interaction of changing weights with continuous registers), the court will assume that such a system is feasible for present purposes.

i. Validity

While it may be possible to develop a weighting system as the plaintiffs' experts described, the evidence shows that such a system would not be substantially equally valid to the defendants' unit-weighting system. First, despite plaintiffs' experts' testimony to the contrary,[36] basing a weighting system on the KSAs underlying an exercise does not make that system more content valid than one using unit weights. As the defendants' experts persuasively testified, such a KSA-based approach suffers from "pseudo-precision." In other words, while the plaintiffs' proposed system appears on the surface to be more content valid than one using unit weights (in that the score is more directly based on the job content as described by the SMEs), it does not actually add any measurable amount of content validity.[37]

Dr. Roth explained this "pseudo-precision" by pointing out that the KSAs identified by the job analysis do not exist as independent entities; rather, they overlap and interrelate with one another. So, under the plaintiffs' weighting scheme, in a job with 20 KSAs, you would need

> "a 20 by 20 matrix of how interrelated are these KSAs. Then you're either counting KSAs or measuring the importance of them and then try[ing] to assign them to particular jobs.... I think you have to realize the inherent limitations to what goes on in those approaches ... [I]t's very hard to
> 1313 say this one KSA is 2.2% of the importance on the job, *1313 therefore we're going to make sure it's 2.2 of the importance on the scoring scheme. I think that's an unrealistic expectation."[38]

These difficulties ensure that any additional content validity gained from taking into account the underlying KSAs would be lost by the imprecision of and inherent difficulty, if not impossibility, in taking into consideration the number of KSAs measured in an exercise, the extent to which they are measured and their relative importance, and producing an exact numerical weight based thereon.

Second, whereas basing a weighting system on the underlying does not make that system more content valid than one using unit weighting, basing a weighting system on the d-statistic actually makes that system less valid than one using unit weighting. Unlike unit weighting and KSA-based weighting, this approach does not appear in the professional literature on different weighting systems,[39] and the plaintiffs have presented no evidence that it is a professionally accepted practice, or to refute the defendants' experts' contention that "any procedure based on inverse values of d .... [is] not good science."[40]

Also unlike the other two weighting methods, a weighting system based on the d-statistic has no justification based on the content of the exercises.[41] Instead, it is based solely on the average differences in scores between two groups, in this case black and white candidates' scores.[42] Without even taking into account the legal appropriateness of taking into account racial differences in exam scores, it is clear that, by being based in part on a factor that has nothing to do with the content of the job (how different racial groups did on the exam), a weighting system that takes into account the d-statistic would be less content valid than one based only on the job's content, such as unit weighting.[43] As Dr. Roth explained about a d-statistic based weighting system,

> "at a conceptual level I think it's going to have a negative impact on the content validity of the exam. Because you're taking factors that are extraneous to the knowledge, skills and abilities that are part of that job and you're suing that to drive the weights. So the greater the impact of that factor on your weighting approach, the more detrimental it is to making a case for the content validity of that exam."[44]

The plaintiffs have failed to establish that their proposed weighting system based on the KSAs underlying an exercise and the exercise's d-statistic is substantially equally valid to the defendants' system of unit weights. As such, they have failed to establish that the defendants are in violation of the consent decree. Paragraph 4(a) requires that the defendants use selection procedures that have been validated in accordance with the Uniform Guidelines, *Reynolds,* 1994 WL 899259, at *8, and the Uniform Guidelines require the defendants to use a
1314 procedure that has lesser adverse impact when it is "substantially *1314 equally valid for a given purpose" to another selection procedure. 29 C.F.R. § 1607.3B. Because the plaintiffs have not established that their weighting system is substantially equally valid to the defendants', they cannot meet their burden of proving the defendants in contempt of the consent decree. *See also* Uniform Guidelines, Question and Answer 52 (stating that, under the Uniform Guidelines, once the user of a procedure has proven the validity of that procedure, "The burden is then on the person challenging the procedure to show that there is another procedure with better or substantially equal validity which will accomplish the same legitimate business purpose with less adverse impact.").

ii. Adverse impact

As an alternative basis for the court's conclusion that the plaintiffs have failed to show the defendants are in contempt of the consent decree, the court finds that the plaintiffs have not proven that the adverse impact of the plaintiffs' weighting system is less than that of the defendants'. Instead, from the evidence presented at trial, it is unclear which weighting system will result in the lower adverse impact.

The first reason it is unclear is because the plaintiffs have produced no evidence about the *actual* adverse impact of either weighting system. While the plaintiffs have produced evidence (in the form of d-statistics) that black candidates have done worse than white candidates on these exams as scored with unit weighting, that evidence does not reflect any difference in the number of blacks and whites eventually hired by the Transportation Department. In other words, as the defendants have argued, the plaintiffs' evidence shows that unit weighting has the potential for adverse impact, not that it has resulted in actual adverse impact.

The court has two reasons for accepting the defendants' distinction between "potential" and "actual" adverse impact. First, both the defendants' and the plaintiffs' experts testified that the Uniform Guidelines define adverse impact in terms of bottom-line selection rates,[45] and Dr. Roth testified that, in his opinion, "a good place to start [calculating adverse impact] would be after hiring would be done, so that you might start the process by looking at

the overall hiring rate for the entire selection system."[46] Again, the plaintiffs' argument in their contempt motions is that the defendants must adopt an alternative weighting system because, under the Uniform Guidelines, "[w]here two or more selection procedures are available which ... are substantially equally valid for a given purpose, the user should use the procedure which has been demonstrated to have the lesser adverse impact." 29 C.F.R. 1607.3B. As such, the court must necessarily use the Uniform Guidelines' definition of adverse impact to address the plaintiffs' argument that the defendants have not conformed with those Guidelines.

Second, the court is compelled to make the distinction between "potential" and "actual" adverse impact in this case because of all the factors that effect the Transportation Department's ultimate hiring decisions but are not reflected by an exam's d-statistic. Specifically, as the defendants repeatedly pointed out, when the Personnel Department is developing a COE, it looks not only at the candidates' exam scores, but also at: whether the
1315 candidate is willing to work in the geographic region in which the job is located; *1315 whether the candidate is willing to do shift work, if required for the available position; whether the candidate is willing to travel overnight, if required; whether the candidate is willing to work part-time/full-time, as needed; and the veterans' preference. And even after the COE is issued, the defendants must take into account the *Frazer/Ballard* no-bypass rule when ultimately filling any positions. Taking all of these factors together, it is quite possible that the defendants ultimate hiring decisions will show no adverse impact, even when the exams at issue do show that impact. In fact, in the case of the Engineering Assistant Examination, exactly that has happened. That exam had a d-statistic of 0.72 (substantially higher than that of the Senor Right-of-Way Specialist (0.41), and very similar to that of the Civil Engineer Manager (0.70) and the Civil Engineer Administrator (0.71) exams at issue here) and yet the adverse impact ratio of the overall selection process for the Engineering Assistant position was 0.954, meaning that the selection rate for blacks was nearly identical to that for non-blacks.[47]

While the plaintiffs disagree with this distinction between "potential" and "actual" adverse impact, their arguments against making a distinction are without merit. In support of their argument that there is no difference between "potential" and "actual" adverse impact, the plaintiffs rely on the Supreme Court's decision in *Connecticut v. Teal,* 457 U.S. 440, 102 S.Ct. 2525, 73 L.Ed.2d 130 (1982). In that case, the Court found that an employer sued for a violation of Title VII may not defend its use of an exam having disparate impact by asserting that the "bottom-line" result of the process achieved the appropriate racial balance. *Id.* at 442, 102 S.Ct. at 2528. The Court found that such an exam, by depriving individuals of employment "opportunities," violated the plain language of Title VII as interpreted by *Griggs v. Duke Power Co.,* 401 U.S. 424, 91 S.Ct. 849, 28 L.Ed.2d 158 (1971).

The plaintiffs argue that, if this court focuses on proof of "actual" adverse impact in hiring, rather than the "potential" adverse impact of the exams at issue, it is allowing the defendants to use the type of "bottom-line" defense prohibited by *Teal.*

The court must reject the plaintiffs' argument, for two reasons. *Teal* is analytically distinguishable from the case at hand. In *Teal,* the exam at issue, which was not shown to be job related, worked as an absolute barrier to further consideration; if a candidate did not pass the exam, he could not move to the next step in the consideration process. 457 U.S. at 443-44, 102 S.Ct. at 2529. The exams in this case, however, do not work on a pass/fail basis; they do not bar anyone from being considered in later stages of the selection process for a position. Instead, these exams are used to rank candidates, with all candidates still eligible for consideration. While the exam in *Teal* worked as an absolute barrier to the next step in the selection process, here much more than the test score can go into whether a person gets the job for which he applied: geographic preference, willingness to travel overnight, willingness to work part-time/full-time, willingness to do shift work, the veterans' preference, the interview, and the application of the *Frazer/Ballard* rule. In theory, a person who had the lowest score on one of the exams at issue could still be hired by the Transportation Department. For this reason, it cannot be said that these exams deprive any individuals of employment opportunities, and *Teal* is therefore inapplicable. *Cf. Teal,*
1316 457 U.S. at 463 n. 8, 102 S.Ct. at 2539 n. 8 (Powell, J, with whom Burger, C.J., and *1316 Rehnquist and O'Connor, JJ., join, dissenting) ("Another possibility is that employers may integrate consideration of test results into one overall hiring decision based on that `factor' *and* additional factors. Such a process would not, even under the Court's reasoning, result in a finding of discrimination on the basis of disparate impact unless the actual hiring decisions had a disparate impact on the minority group.") (emphasis in original).

Moreover, any other reading of *Teal* would not allow any principled application of its rejection of the "bottom line." There would be no principled way to determine where, or at what point, in a multi-step selection process, adverse impact should be determined. Indeed, once *Teal* is unhinged from the above limited reading, there is no principled reason why adverse impact should not be determined for each and every question separately in an exam, if not each step in the development of a question, which would, of course, be impractical.

The second reason the court must reject the plaintiffs' *Teal* argument is because there is no Title VII issue before the court. This matter is before the court on the plaintiffs' contempt motions; the question is whether the defendants have violated the parties' consent decree by failing to abide by the Uniform Guidelines. The defendants' "potential" versus "actual" adverse-impact argument is not asserted as a "bottom-line" justification of a Title VII violation; rather, it is an argument for why the plaintiffs have not shown the defendants to be in violation of the Uniform Guidelines. For that reason, the plaintiffs' reliance on *Teal,* a case that applies § 703(a)(2) of Title VII, is entirely misplaced. For purposes of plaintiffs' contempt motions, the Uniform Guidelines definition of adverse impact is controlling. Using that definition, which looks at adverse impact in terms of bottom-line selection rates, it is unclear which weighting system results in lesser adverse impact; the plaintiffs have produced no evidence as to the actual adverse impact of unit weighting versus another weighting system, and the plaintiffs have therefore failed to meet their burden of proving that the adverse impact of the plaintiffs' weighting system is less than that of the defendants'.

The second reason it is unclear whether the plaintiffs' or the defendants' weighting system will result in lesser adverse impact is because, while the plaintiffs have provided some evidence that the potential adverse impact of using unit weighting is high, they have not proved that the potential adverse impact of their system is lower.

It is undisputed that black candidates' scores under unit weighting are worse than white candidates' scores under that weighting system. The d-statistics of all five exams are positive, in this case meaning that blacks did worse than non-blacks on all exams, and range from 0.41 to 1.16. Dr. Lefkowitz testified that the 1.16 composite d-statistic for the Civil Engineer-Design Option Examination is very large, "substantially larger even than where we're used to seeing on the sorts of paper and pencil tests of general mental abilities that are known to produce the large kind of adverse impact. This is even larger than that."[48] Furthermore, only candidates who passed the MQ requirements were allowed to take these exams, which is significant because, as shown in an article by the defendants' experts, "you will get an underestimate of the adverse impact of an examination in a multiple hurdle
1317 situation in which people have been pre-screened previously, especially if there is *1317 adverse impact on the previous screening procedure."[49]

On the other hand, it is unclear how reliable these d-statistics are. The "substantially large[]" 1.16 d-statistic was not statistically significant at the 5% level; indeed, the d-statistics on only two of the five tests were statistically significant, and there was some evidence presented that the difference in black candidates' scores and white candidates' scores could be due to sample size as much as any true difference between how blacks and whites could be expected to score on these exams. Further, the defendants took a number of measures to minimize the adverse impact of these exams, including: (1) the choice to use work-sample exams, which the plaintiffs' experts admitted have less adverse impact than traditional paper and pencil measures;[50] (2) having the exercises reviewed and edited by an external psychologist; (3) having black and female incumbents serve as SMEs; (4) ensuring that no exercise was rated as biased; (5) putting at least one black and one female rater on each assessment panel; (6) giving a how-to-prepare manual to each candidate; and (7) having the plaintiffs' and defendants' experts review the exams at several stages in the development process.[51] While these two factors (the lack of statistical significance of three exams' d-statistics, and the measures taken by the defendants to minimize adverse impact) do not excuse the fact that black candidates scored lower than white candidates on these exams, they do make it harder for the plaintiffs to meet their burden of proving that their weighting system would have less potential adverse impact than the defendants' system.

Most importantly, the plaintiffs have not shown by clear and convincing evidence that their weighting system will have lower potential adverse impact that the defendants' weighting system. First, the plaintiffs' experts admitted that they have not run the numbers to determine which system has lower adverse impact.[52] While they think it

is likely that the plaintiffs' weighting system would result in a lower d-statistic, the plaintiffs' experts don't know that for a fact.[53] In fact, one of the defendants' experts, Dr. Buster, testified that, while he had not calculated the numbers, one could argue that unit weighting was more favorable to adverse impact than the plaintiffs' proposed system.[54]

Taking the experts' testimony together, the court finds that the plaintiffs have failed to meet their burden of showing that their weighting system will have lower potential adverse impact than the defendants' weighting system. While the plaintiffs' system would rely in part on the d-statistic (presumably lowering the potential adverse impact of their weighting system as compared with unit weighting), it would also rely on the underlying KSAs. The plaintiffs, however, produced no evidence about the effect that building a weighting system on the underlying KSAs would have on the potential adverse impact of that system. Because the plaintiffs have not done the calculations, the court cannot assume, simply because it relies in part on the d-statistic, that the plaintiffs' weighting system would result in lower potential adverse impact than the defendants' system. As such, for this
1318 reason as well, the plaintiffs have failed to meet their burden of proving that the adverse impact of the *1318
plaintiffs' weighting system is less than that of the defendants'.

Finally, because the court has found that the plaintiffs have not proven that their weighting system has less adverse impact than the defendants' system, the court need not address the parties' dispute over the appropriateness of basing a weighting system on the d-statistic, insofar as it measures the differences in test scores based upon the test-taker's race. *Compare* Hayden v. County of Nassau, 180 F.3d 42, 48 (2d Cir.1999) ("The only manner in which race was implicated is that Nassau County set out to design an entrance exam which would diminish the adverse impact on black applicants. This desire, in and of itself, however, does not constitute a `racial classification.' Since the exam was administered in a race-neutral fashion which did not expressly distinguish between applicants on the basis of race, Nassau County's intent, without anything more, does not implicate an express, racial classification."); *with* San Francisco Police Officers' Ass'n v. City of San Francisco, 869 F.2d 1182, 1184 (9th Cir.1988) ("When, however, the City arbitrarily changed the weighting [to increase minority candidates' scores] and promoted on the basis of this weighting, the City violated the Consent Decree by discriminating on the basis of race and sex. The victims of this discrimination were . . . the larger group of all the white males who took the test and who were passed over by the discriminatory weighting and subsequent promotions.").

b. Banded scoring

The plaintiffs' second argument for finding the defendants in contempt of the consent decree is that these tests are not sufficiently valid to be used for the purpose of rank-ordering candidates' scores. Instead, the plaintiffs propose that the defendants use banded scoring, whereby similar exam scores would be placed in scoring bands, rather than ordinally ranked.

1. Waiver

The defendants' first response to the plaintiffs' banding argument is that the plaintiffs waived this argument by failing to raise it either in their contempt motions or in their expert report.

In a previous appeal in this case, the Eleventh Circuit explained how consent decrees are to be enforced:

> "If the plaintiff (the party obtaining the writ) believes that the defendant (the enjoined party) is failing to comply with the decree's mandate, the plaintiff moves the court to issue an order to show cause why the defendant should not be adjudged in civil contempt and sanctioned. The plaintiff's motion cites the injunctive provision at issue and alleges that the defendant has refused to obey its mandate. If satisfied that the plaintiff's motion states a case of non-compliance, the court orders the defendant to show cause why he should not be held in contempt and schedules a hearing for that purpose. At the hearing, if the plaintiff proves what he has alleged in his motion for an order to show cause, the court hears from the defendant. At the end of the day, the court determines

whether the defendant has complied with the injunctive provision at issue and, if not, the sanction (s) necessary to ensure compliance."

Reynolds v. Roberts, 207 F.3d 1288, 1298 (11th Cir.2000) (citations omitted). The defendants are correct that, in their contempt motions, the plaintiffs did not allege the defendants' failure to adopt banded scoring as one of their bases for finding the defendants in contempt of the consent decree. On the other hand, the plaintiffs did allege in their motions that the defendants are in violation of ¶ 4(a) of Article Three of the consent decree, and therefore
1319 complied with the above language, which required that their motions for contempt *1319 "cite[] the injunctive provision at issue and allege[] that the defendant has refused to obey its mandate." *Id.* Although the court does not approve of the plaintiffs' apparent last-minute addition of their banded-scoring claim, they do appear to have compiled with the above language, and, more importantly, the defendants do not appear to have been prejudiced by the plaintiffs' introduction of this claim. As such, the court will consider the plaintiffs' banded-coring claim.

2. Validity of rank ordering

While the plaintiffs have objected to the defendants' rank ordering of candidates' test scores, and have proposed that the defendants use banded scoring instead, they have not undertaken to show that banded scoring is as content valid as ranking, or that it would have less adverse impact than ranking.[55] Instead, the plaintiffs focused on proving that the defendants had not shown their exams to be sufficiently valid to be used on a ranking basis.[56] As such, it does not appear that the plaintiffs are arguing that the defendants must use banded scoring because it is a procedure that has less adverse impact than ranking and is "substantially equally valid for a given purpose." 29 C.F.R. § 1607.3B. Instead, the court assumes that the plaintiffs' argument is that the defendants are not in compliance with ¶ 4(a) of Article III of the consent decree because the defendants have not complied with the Uniform Guidelines' requirement that they show these exams are sufficiently valid to be used on a ranking basis. 29 C.F.R. § 1607.14C(9) ("If a user can show, by a job analysis or otherwise, that a higher score on a content valid selection procedure is likely to result in better job performance, the results may be used to rank persons who score above minimum levels.").

Unlike above, where the burden was on the plaintiffs to show that their weighting system was substantially equally valid to the defendants, here the Uniform Guidelines clearly put the burden on the defendants to show that these exams are sufficiently valid to be used for ranking. *Id.* The court has already found that these five exams are highly content valid, so the defendants' remaining burden is to show "that a higher score on a content valid selection procedure is likely to result in better job performance." *Id.* Whether there has been a sufficient demonstration that an exam may be used on a ranking basis is a matter that is within the bounds of acceptable professional practice,[57] and it is within the professional judgment of the developer as to how to show that ranking is appropriate.[58]

To meet their burden, the defendants have provided extensive expert testimony that their exams may be used on a ranking basis. First, Dr. Bobko testified that, in his opinion, "because these exams are content valid exams and they show the kinds of properties that you would want them to show and the way they were built out ... that it is
1320 appropriate to use these in a ranking fashion."[59] As an example, Dr. Bobko pointed out that, because the KSAs *1320 were chosen based on their relation to acceptable job performance, these exams had the right fundamental building blocks to predict that higher scores on these exams would lead to better job performance.[60] Dr. Bobko also noted that the SMEs were asked how adequately each exercise distinguishes between superior and adequate levels of performance, which "confirm[s] that the kinds of exercises that had been built were, indeed, related to job performance."[61] Finally, Dr. Bobko also explained that the use of critical incidents to develop the test exercises makes it likely that a higher score on these exams would lead to better job performance.[62] Dr. Buster agreed that, on average, a person with a higher score on these exams will, on average, exhibit better job performance on average.[63]

Additionally, Dr. Bobko testified that the professional literature in his field supports the defendants' use of these exams to rank candidates' scores. Dr. Bobko noted that a set of principles for the Society of Industrial Organizational Psychology indicate "that if an exam has been demonstrated to be content valid and there is an

adequate range of variation of the scores, that one can use ranking and presume that higher scores are associated with higher performance."[64] Quoting these principles, Dr. Bobko testified that "selection techniques developed by content-oriented procedures and discriminating adequately within the range of interest can be assumed to have a linear relationship to job behavior. Consequently, ranking on the basis of such scores is appropriate."[65] Dr. Bobko believes that these exams exhibit an adequate range of scores, and therefore can be used to rank candidates.[66]

In response, the plaintiffs' experts testified without elaboration that they did not believe the exams at issue have been shown to be sufficiently content valid to be used on a ranking basis.[67] Dr. Outtz admitted that, in theory, a higher score on these exams might lead to better performance, but said that the performance differences you would predict from small differences in scores are too small to detect. "So that if one person scores a fourth of a point higher than the other, true enough in theory, they would have a predicted difference in performance. The actual difference that you would find would be minuscule and inconsequential."[68]

The defendants experts, however, responded extensively to Dr. Outtz's criticism. Dr. Roth testified that, while a very small difference in scores may not indicate that the candidate with the higher exam score will perform better on the job, "across a large number of employment decisions in a case where you are only able to give a test to each individual at one point in time, that hiring the highest scoring individual will on average give the state the
1321 best possible workers that it can."[69] Similarly, Dr. Bobko testified that, while it cannot be assumed that a *1321 difference as small as a fourth of a point one these exams will indicate better job performance, "if one person scores higher than another person, that is the best available evidence I have to predict from a content valid test that the first person will outperform the second person by average. That's the best available evidence I have. One person scores higher on a test, if the test is built out in a content valid way I would predict that person will also perform better."[70]

Despite the plaintiffs' criticisms, the court finds that the defendants have shown that a higher score on these exams "is likely to result in better job performance," 29 C.F.R. § 1607.14C(9), and may therefore rank candidates based upon their scores on these exams. These exams are highly content valid reflecting quite closely the content of the underlying jobs and the SMEs have evaluated the exam exercises to ensure that they distinguish between different levels of job performance. Furthermore, Dr. Bobko testified that there is an adequate variation in exam scores such that, under principles for the Society of Industrial Organizational Psychology, one can use ranking and "presume that higher scores are associated with higher performance."[71] As such, the defendants have met their burden of showing that a candidate who has a higher score on these exams is likely to exhibit better job performance; they are therefore in compliance with the Uniform Guidelines and cannot be found to be in contempt of ¶ 4(a) of Article III for ranking candidates by score.

[1] The Adams intervenors, who represent non-African-Americans, joined the defendants in opposing the plaintiffs' motions. For present purposes, there is no need to distinguish the defendants' and Adams intervenors' arguments.

[2] Tr. of May 8, 2003, at 70 (testimony of plaintiffs' expert Dr. Joel Lefkowtiz).

[3] For three of the exams at issue (Civil Engineer Design, Civil Engineer-Construction Option, and Senior Right-of-way Specialist), the SMEs were allowed to give only one response to this question. For the other two exams they were allowed multiple responses.

[4] On the Civil Engineer Manager and Civil Engineer Administrator Examinations, the SMEs were asked, "To what extent does a high overall score on this exercise predict how well a candidate will perform on the job? (0 = Not Likely, 1 = Likely, 2 = Very Likely)." *See, e.g.,* defs.' ex. 7 (Content Validation Report: Civil Engineer Manager), at 11.

[5] Performance dimensions are groups of KSAs. The KSAs were grouped by the SMEs into target areas of performance based on the general focus or meaning of the KSA. The SMEs were not asked to answer this question for the Civil Engineer Manager and Civil Engineer Administrator Examinations.

[6] *See, e.g.,* defs.' ex. 6 (Content Validation Report: Civil Engineer-Construction Option), at 20.

[7] *See* report and recommendation of Special Master Carlos González, entered March 28, 2003 (Doc. no. 6537) (discussing purpose for and process of developing MQs).

[8] Tr. of May 8, 2003, at 17-18.

[9] Tr. of May 9, 2003, at 97, 99.

[10] *See, e.g.,* tr. of May 8, 2003, at 62.

[11] If an exercise had multiple components, the score for each component was standardized separately and then weighted using unit weighting. For example, for the Civil Engineer Construction Option exam, the role play exercise was scored for both content and oral communication, and the payment voucher exercise was scored for content and written communication. Thus, while this exam only had four exercises, it resulted in six scores, each of which was standardized, unit weighted, and added together to get one final score.

[12] Depending on a veteran's status, he will be given a five- or ten-point addition to his score. This is done pursuant to state law.

[13] The *Frazer/Ballard* no-bypass rule states:

"Defendants shall not appoint or offer a position to a lower-ranking white applicant on a certificate in preference to a higher-ranking available Negro applicant, unless the defendants have first contacted and interviewed the higher-ranking Negro applicant and have determined that the Negro applicant cannot perform the functions of the position, is otherwise unfit for it, or is unavailable. In every instance where a determination is made that the Negro applicant is unfit or unavailable, documentary evidence shall be maintained by the defendants that will sustain that finding." *United States v. Frazer,* 317 F.Supp. 1079, 1091 (M.D.Ala.1970).

[14] Report and recommendation, entered March 28, 2003 (Doc. no. 6537); order entered May 9, 2003 (Doc. no. 6719).

[15] The plaintiffs' experts, as stated, are Drs. Lefkowitz and Outtz, both of whom testified at the hearing held on this matter. The defendants' experts are Dr. Philip Bobko and Dr. Philip Roth, who testified at the hearing, and Dr. Maury Buster, who testified by deposition.

[16] Tr. of May 8, 2003, at 70.

[17] *See also* defs.' ex. 4 (Defendants' Experts' Report), at 15.

[18] Tr. of May 9, 2003, at 143.

[19] Tr. of May 8, 2003, at 46.

[20] Tr. of May 9, 2003, at 132-33.

[21] The plaintiffs also object to the defendants' choice of a 50%, rather than 67%, cutoff for the necessary-at-entry screen. The court rejects this objection as well. As the special master found, "[t]here is ample evidence in the record to support the [defendants'] decision to use ... a .50 linkage screen." Report and recommendation of Special Master González, at 63.

[22] Tr. of May 9, 2003, at 131.

[23] Report and recommendation, at 61-62.

[24] At the hearing on this matter, the plaintiffs presented three alternative weighting plans for the Civil Engineer Construction Option Examination. The first of these plans assigned weights based on a simple count of the number of KSAs measured by each exercise; the second plan was a variation on the first that took into account the fact that the two technical exercises measured some of the same KSAs; and the third was a further

variation on the second that took into account the fact that both the technical and non-technical exercises measured the same KSAs. Pls.' exs. 461, 462 and 463; *see also* tr. of May 9, 2002, at 51-57. The plaintiffs' experts appeared to concede, however, that these plans were simply examples, and that a rating system based on the KSAs underlying an exercise would also look at how important each KSA was and to what extent it was measured by an exercise. Tr. of May 8, 2003, at 20-23.

[25] Defs.' ex. 4, at 15.

[26] Deposition of Dr. Maury Buster, at 88.

[27] *Id.* at 16.

[28] Tr. of May 8, 2003, at 117.

[29] *Id.* at 118.

[30] *Id.* at 122; tr. of May 9, 2003, at 55.

[31] Tr. of May 8, 2003, at 119, 143.

[32] Deposition of Dr. Maury Buster, at 16-18.

[33] Notably, the plaintiffs are not arguing that the defendants should look for and use an "alternative device[]." Instead, they are arguing that the defendants should make alternative *use* of the devices at issue—the exams—by weighting the scores differently. Paragraph 8 nowhere requires the defendants to consider alternative uses of a device. *Compare* 29 C.F.R. § 1607.3B ("[W]henever a validity study is called for by these guidelines, the user should include ... an investigation of suitable alternative selection procedures *and* suitable alternative methods of using the selection procedure . . .") (emphasis added).

[34] The plaintiffs also assert in their motions for contempt relief that the defendants have not shown that unit weighting and ranking are a business necessity. The plaintiffs, however, have not argued to the court how this failure violates the consent decree. *Cf. Hamer v. City of Atlanta,* 872 F.2d 1521, 1534 (11th Cir.1989) ("Since the district court and this panel find that the test was properly validated [pursuant to the Uniform Guidelines], and since the appellants have as their only appellate contention that the test was nor properly validated, we conclude that there is no error in the district court not having considered this [business necessity] factor."). Because this matter is before the court on the question of whether the defendants are in contempt of the provisions of the parties' consent decree, the plaintiffs' failure to assert a violation of that decree requires the court to find the plaintiffs' business-necessity argument to be procedurally improper. *Cf. Reynolds v. Roberts,* 207 F.3d 1288, 1298 (11th Cir.2000) ("The reason why plaintiffs' counsel did not move the court for an order to show cause is obvious: the Department had not disobeyed any of the mandates of the consent decree, as amended, and plaintiffs' counsel could not contend that it had without running afoul of Rule 11 of the Federal Rules of Civil Procedure.").

[35] *See, e.g.,* tr. of May 8, 2003, at 24-25; tr. of May 9, 2003, at 58-59.

[36] *See, e.g.,* tr. of May 9, 2003, at 61.

[37] *See* tr. of May 8, 2003, at 156 (stating that defendants' weighting system is just as content valid as plaintiffs'); *see also id.* at 96 ("[Plaintiffs' experts] never had any a priori objection to unit weighting, nor do we have any focused objection on unit weighting now. The point is that unit weighting does not contribute to the content validity and consequently is no more valid than other weighting schemes. We never voiced any objection to unit weighting, per se.").

[38] Tr. of May 9, 2003, at 159-160; *see also* tr. of May 8, 2003, at 156 (noting KSA-based weighting system is "a misguided notion" and is "unnecessarily complex").

[39] Tr. of May 9, 2003, at 156.

[40] Defs.' ex. 4, at 17; *see also* tr. of May 8, 2003, at 51 (stating that Dr. Lefkowitz does not know of any instances where a d-statistic based weighting scheme has been used).

[41] Tr. of May 8, 2003, at 127; tr. of May 9, 2003, at 153.

[42] *Id.*

[43] *Id.* at 130; defs.' ex. 4, at 17.

[44] Tr. of May 9, 2003, at 154.

[45] *Id.* at 84, 134.

[46] *Id.* at 134; *see also* 29 C.F.R. § 1607.16 (definitions) ("*Adverse impact.* A substantially different rate of selection in hiring, promotion, or other employment decision which works to the disadvantage of members of a race, sex, or ethnic group.").

[47] Defs.' ex. 4, at 11.

[48] Tr. of May 7, 2003, at 193-94.

[49] *Id.* at 194-95.

[50] Tr. of May 8, 2003, at 39.

[51] *See, e.g.,* defs.' ex. 6, at 35.

[52] Tr. of May 8, 2003, at 25.

[53] Tr. of May 9, 2003, at 83.

[54] Deposition of Dr. Maury Buster, at 22-23.

[55] In fact, only the defendants presented evidence about banded scoring, and their evidence tended to discredit its use. *See, e.g.,* deposition of Dr. Maury Buster, at 95-97 (stating that bands include too many scores the standard error of measurement is higher for lower scores than for higher scores, but banding has its greatest effect on the top scores).

[56] *See, e.g.,* tr. of May 8, 2003, at 66 (asking plaintiffs expert only if exams at issue are sufficiently valid to be used on a ranking basis).

[57] Tr. of May 8, 2003, at 66.

[58] *Id.* at 33.

[59] *Id.* at 109.

[60] *Id.* at 110.

[61] *Id.* at 112.

[62] *Id.* at 111.

[63] Deposition of Dr. Maury Buster, at 102.

[64] Tr. of May 8, 2003, at 113; *see also* defs.' ex. 27 ("Principles for the Validation and Use of Personnel Selection Devices" (SIOP 1987)).

[65] Tr. of May 8, 2003, at 115.

[66] *Id.*

[67] *Id.* at 66; tr. of May 9, 2003, at 71.

[68] Tr. of May 9, 2003, at 72.

[69] *Id.* at 187.

[70] Tr. of May 8, 2003, at 173.

[71] *Id.* at 113.