EDTPA:
BIAS AND RELATED EQUITY ISSUES

OCCASIONAL REPORT NO. 13.1
ASSESSMENT & ACCREDITATION COMMITTEE

EDTPA:
BIAS AND RELATED EQUITY ISSUES

A Take on the Pilot Data

During the edTPA-Tk20 professional development session faculty members raised pertinent questions about racial/ethnic bias and other equity issues associated with the edTPA. We have located limited information about these issues—most based on the '12-'13 pilot study. Please note that a reference section is included at the end of this brief paper.

While no gender difference accrued, scores differed by racial and ethnic group. Though these score differences proved very small, they would probably be statistically significant—if for no other reason than sample size ~ 3669; edTPA, 2013). It remains perhaps a bit disappointing that these inferential tests were not run, either in the spring 2013 field test or during the Pecheone and Chung Wei (2007) Performance Assessment for California Teachers (PACT) review.

The PACT instrument served as the immediate predecessor to the TPA, that, in turn, evolved into the edTPA. Darling-Hammond (2013) traces both PACT and TPA to the National Board for Professional Teaching Standards (NBPTS) master teacher portfolio.

As it turns out, a bias committee exists that will look closely at the edTPA re these issues. It remains important to identify the membership of this committee—perhaps, as one of the more sizable participants, SCSU (or MACTE) ought to request membership [on this body].

The variables used in this brief analysis are total scores, running from 15 to 75. This produces an overall mean score of roughly 43 (see Table 1).

We estimated effect sizes for each white-other comparison and, as advertised, they proved quite small. These differences are smaller than characteristically seen on paper-pencil or electronic tests (typically a z of ~ 1.0). These estimates are quite rough in that we averaged (weighted for sample size) instead of pooled, SDs—but that usually gets us pretty close to the true effect Size parameter. In order to pool the SDs, we would need access to the original data. Right now, a joint sample from MnSCU institutions is being developed that may be large enough for making such parameter estimates, but this remains a work in progress.

The effect size (ES) for Black v white candidates equaled .36, favoring white candidates; this translates to about a third of a standard deviation, as these estimates are roughly equivalent to z scores and thus can be interpreted as standard deviation units. The ES for Hispanic (SCALE language) vs. white candidates = .053, very slightly favoring white candidates, almost certainly a non-significant or at least uninterpretable difference.

A Tentative Conclusion re Ethnic Differences

During the national field test, a small, but significant Black vs. white difference was observed. This difference amounted to about 1/3 of an SD—still WAY too much--but lower than the differences typically observed on traditional teacher tests. Good news and bad news.

Table 1. Summary statistics for total score by subgroups (reproduced from the edTPA pilot study (2013)

Summary Statistics of Total Score by Subgroups			
Gender	**N**	**Mean**	**SD**
Female	2,819	42.88	8.25
Male	795	42.23	7.89
Undeclared	55	44.55	7.65
Ethnicity	**N**	**Mean**	**SD**
American Indian or Alaskan	19	39.74	9.60
Asian or Pacific Islander	145	44.97	8.09
Black (Non-Hispanic)	83	39.67	8.68
Hispanic	143	42.28	6.85
Other	90	42.58	9.27
Uncategorized	184	44.36	8.11
White (Non-Hispanic)	3,005	42.69	8.14
Primary Language	**N**	**Mean**	**SD**
Non-English	62	42.29	8.07
English	3,568	42.75	8.16
Undeclared	39	44.74	9.55

Differences between Rubrics and Areas

It is a little alarming that systematic differences were observed across the rubrics themselves. The lowest was Rubric 13 (Assessment: Student Use of Feedback) @ 2.38, with the highest being Rubric 1 (Planning: Planning for Subject-Specific Understandings) @ 3.15. Also, Task 3 (Assessment) proved significantly lower than did the other two tasks. We hope that developers will adjust the manuals. Rubric-based discrepancies raise the following questions:

1. Does the d resonate with instrumental unreliability or does it reflect real (e.g., reliable and valid) differences?

2. Is reliable between-rubric variance (if any exists) associated with difficulties completing tasks correctly, with shortcomings in preparation, or (most probably) with some interaction between these factors.

Perhaps it is time that we complete a qualitative investigation by deprogramming some of our candidates who scored at each level of the assessment rubrics. Let's ***ask*** them!

Pertinent Summary of the Issues Surrounding Validity by Group

The following summary statement (Lam, 1995) struck us as a useful summary of what we will seek in demonstrating the reliability, validity, sensitivity, and equity of the edTPA ***before*** it becomes a high stakes test for our candidates. The numbered list could reasonable serve as a framework for research studies to be completed on the edTPA—both nationally and locally.

> Traditional tests with selection response items have been criticized as unfair to minority students because these students typically perform less well on this type of test than majority students. However, no evidence is yet available to substantiate the claim that performance assessment can in fact diminish differential performance between groups (Linn et. al., 1991). Although the use of performance assessment can eliminate some sources of bias, such as testwiseness in selecting answers that are

associated with traditional tests, it fails to eliminate others, such as language proficiency, prior knowledge and experience, and it introduces new potential sources of bias:

1) ability to handle complex problems and tasks that demand higher order thinking skills (Baker& O'Neil, 1993);

2) metacognitive skills in conducting self-evaluation, monitoring thinking, and preparing and presenting work with respect to evaluation criteria;

3) culturally influenced processes in solving problems (Hambleton & Murphy, 1992);

4) culturally enriched authentic tasks;

5) low social skills and introverted personality;

6) added communication skills to present, discuss, argue, debate, and verbalize thoughts;

7) inadequate or undue assistance from parents, peers, and teachers;

8) lack of resources inside and outside of schools;

9) incompatibility in language and culture between assessors and students; and

10) subjectivity in rating and informal observations.

A strategy for reducing the influence of extraneous factors in rating that also supports integration of curricula is to employ multiple scales for different attributes embedded in the performance. For example, essays on social studies can be rated on subject matter knowledge, writing quality, and penmanship. (Lam, 1995)

References

Au, W. (2013) What's a nice test like you doing in a place like this? The edTPA and corporate reform. *Rethinking Schools*, X (X), 24-27.

Darling-Hammond, L., & Hyler, M. E. (2013). The role-performance assessment in developing teaching as a profession. *Rethinking Schools,* X(XX), 10-15.

edTPA (2013). 2013 edTPA field test: Summary report. Palo Alto, CA: Stanford Center for Assessment, Learning, and Equity.

Gorlewski, J. (2013). What is the edTPA and why do critics dislike it? Entry on Diane Ravitch's blog: http://dianeravitch.net/2013/06/03/what-is-edtpa-and-why-do-critics-dislike-it/ (downloaded on 11/6/13).

Hilburn, J. (2011). Letter from the Editorial Board. *High School Journal*, *94*(3), 79-81.

Lam, T. (1995). *Fairness and equity in performance assessments*. Available from ERIC Digest (ED391982): http://www.ericdigests.org/1996-4/fairness.htm/

Madeloni, B., & Gorlewski, J. (2013). Wrong answer to the wrong question. *Rethinking Schools*, *X*(X), 16-22.

NCATE & SCALE (2012, Oct. 22). Statement on how the edTPA can be used to meet NCATE standards. http://edtpa.aacte.org/wp-content/uploads/2013/01/edTPA-QA-about-NCATE-Accreditation-10-12-20121.pdf, downloaded on November 10, 2013.

Pecheone, R.L., & Chung Wei, R. R. (2007). PACT Technical Report: Summary of Validity and Reliability Studies for the 2003-04 Pilot Year. Palo Alto, CA: Stanford University. Downloaded on November 10, 2013 from http://www.pacttpa.org/_files/Publications_and_Presentations/PACT_Technical_Report_March07.pdf. (See section starting on page 27.)

Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education IN TEACHER EDUCATION: THE PERFORMANCE ASSESSMENT FOR CALIFORNIA TEACHERS (PACT). *Journal Of Teacher Education*, *57*(1), 22-36. doi:10.1177/0022487105284045

Sandholtz, J., & Shea, L. M. (2012). Predicting performance: A comparison of university supervisors' predictions and teacher candidates' scores on a teaching performance assessment. *Journal Of Teacher Education*, *63*(1), 39-50. doi:10.1177/0022487111421175