Score Comparability Study of Online and Paper-Pencil Administrations Using Propensity Score Matching Models

Dong Gi Seo

Michigan Department of Education

## Objectives of the Inquiry

State education departments are rapidly exploring and adapting online assessments as part of their statewide assessment programs. As part of an ongoing plan to transition to online testing, many states are offering an online pilot test to a limited number of schools that have had the infrastructure and equipment to test online. For this reason, an online test was implemented side by side with a paper-pencil test. Thus, professional testing standards should be considered to ensure comparable results across paper and online modes.

The theoretical assumptions underlying procedures for establishing score comparability have been sufficiently satisfied. Standard 4.10 in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999)

states that: “Support should be provided for any assertion that scores obtained using different items or testing materials, or different testing procedures, are interchangeable for some purpose. This standard applies, for example, to alternate forms of a paper-and-pencil test or to alternate sets of items taken by different examinees in computerized adaptive testing” (p.57).

The comparability of test scores between online and paper version has been studied by many researchers (e.g., Ito & Sykes, 2004; Paek, 2005, Poggio, Glassnapp, Yang, & Poggio, 2005; Pommerich, 2004). In general, the results of comparability research showed that online and paper versions are comparable across grades and academic subjects. However, a few comparability studies have seriously considered the procedure of matching the online-based sample with the paper-based sample (e.g., Way, David, & Fitzpatrick, 2006; Karkee, Kim, & Fatica, 2010).

The purpose of this study is to evaluate the comparability of a paper and online version of a statewide assessment for the purposes of test score reporting, and to appropriately adjust equated score conversion tables for students testing online as warranted. In the sections that follow, this study will describe initial efforts to transition the program to online testing, and introduce the design and methodology used for the comparability studies at each grade level, and present results of the score comparability studies conducted at grade 6 and 9 social studies.

In particular, this study will introduce an approach and design to study the comparability of online and paper tests with a propensity score matching method. The propensity score is used to match the sample of online tests with the sample of paper pencil tests. Finally, we will report on some additional analyses that evaluate the sensitivity of the propensity score matching approach for diminishing differences in online and paper group performance when these groups differ in terms of overall proficiency.

## Method

### Data Resources

Data from a high-stakes social studies assessment for Grade 6 and Grade 9 in a Midwestern state of the US were used to investigate scale comparability. Table 1 describes the demographic characteristics of the students.

---

Insert Table 1

---

In this social studies test, each form consists of operational and field test items. Five forms were created for the paper test and one of those was chosen for the online social studies pilot. The states' assessments are given in the fall and measure content from the previous grade.

All these forms were built under the same test specification (blueprint) that mapped the state social studies curricula and standards. Table 2 and 3 described social studies Grade

6 and 9 test blueprints reflecting five major social studies areas. Within each grade, the forms are parallel and test scores were post equated.

---

Insert Table 2

---

---

Insert Table 3

---

**Matching Variables**

Students with missing values in matching variables are excluded in this study. This study used the student demographic information from the dataset provided by in the student data system, which includes gender, ethnicity, native language, limited English proficiency, special education status, and economically disadvantaged status. Student achievement in the current school year's social studies assessment was obtained from student level achievement datasets as an internal matching because the addition of the external matching variables (e.g., scale scores in the other content areas) do not add significant information (Karkee, Kim, & Fatica, 2010).

For school level variables, some variables in the student level (e.g., number of females and number of students) were aggregated at the school level. In addition, information on the number of students per computer and online speed were also considered to match variables at the school level.

**Propensity Score Matching (PSM)**

PSM is designed for causal inference with a dichotomous treatment variable and a set of pretreatment control variables, X. If the treatment variable is binary, denoted by Z=1 versus Z=0, the conditional probability of assigning each experimental unit to a particular treatment can be written as an a priori function of X as follows,

$$e(X) = \mathrm{P}(Z = 1 \mid X) \qquad (1)$$

Where $e(X)$ is called the propensity score. Rosenbaum and Rubin (1983) proved that the treatment assignment and the treatment variables are conditionally independent given the propensity score regardless of how X is distributed. Under this assumption, the observed outcome of a control group provides an unbiased estimate of the counterfactual outcome of a treatment group when two groups are drawn from the same joint distribution of all the observed variables X.

MATCHIT package (Ho, Imai, King, and Stuart, 2007) was used to implement a matching data set, which works in conjunction with the R program (R Development Core Team, 2008). An optimal pair-matching approach was used (by using MATCHIT library in R) in this study. "Optimal" matching finds matched samples with the smallest average absolute distance across all matched pairs. Gu and Rosenbaum (1993) found that optimal matching does a better job of minimizing the distance within each pair and is helpful when there are not many appropriate control matches for the treatment group.

Balance statistics (e.g., t-test) were then used to test all variable balances (student level variables at the student level and school level variables at the school level) between the matched paper-pencil students and the online students. The main focus was given to achievement related variables at both the student and school levels.

**Equating Scale Scores**

*Equating separately,* WINSTEPS and a fixed parameter approach was used to calibrate items' parameters for the online group and the matched paper-pencil group respectively. The estimated item parameters were equated to the base scale and, and the conversion tables for the online group and the matched paper-pencil group were created respectively.

*Equating together,* WINSTEP and a fixed parameter approach were used to calibrate items' parameters for the combined online and the matched paper-pencil group data. The estimated item parameters were equated to the base scale, and then the single conversion table was created for both online and paper-pencil group.

**Comparability Analyses**

*Differential item functioning and differential test functioning (DIF & DTF).* In order to review the favor group, DIF and DTF were reviewed for online and paper-pencil students.

*Mean difference of latent trait scale* the latent trait scale scores of online students were compared with those of the matched paper-pencil students

*Mean difference of scale scores* the scale scores of online students using the online conversion table were compared with those of online students using the conversion table from all students.

*A difference of performance level* chi-square test was used to test any notable difference between students who are proficient in the online conversion table and those in the conversion table from all students.

## Preliminary Results

Variables at student levels were examined for a balance check. Everything else at student level did not show significant differences between two groups. Table 4 showed mean and standard deviation (SD) of latent trait and scale scores for the both online and matched paper-pencil sample. There were no significant mean differences of latent trait and scale scores between the online and matched paper-pencil samples.

For a performance level test, a few students taking an online test were identified as not proficient if the conversion table obtained from all students was used even if they were classified as a proficient from the conversion table based on their own data. However, chi-square test did not show a significant effect. Therefore, a decision of combining all students was made to provide students with their scale scores, and the online students were treated as if they were administered under the paper-pencil test. After the overall conversion table was made for each assessment level, student performance levels were not influenced by a mode effect.

## Educational Importance of the Study

The purpose of this study is to bring justifiable attention to the issues that no special provisions appear necessary to offer simultaneously both online and paper-pencil assessments. Based on this study, the separated assessment coupled with the need for separated conversion table would not be required in a state's assessment. Through a score comparability study, dual programs coupled with the separated conversion table do not appear needed, and additional psychometric manipulation appears unnecessary. As a result, the online test appears to provide a credible and comparable option to the paper-pencil modality.

## Reference


American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME). (1999). Standards for educational and psychological testing. Washington, DC: AERA.

Glassnapp, D. R., Poggio, J., Poggio, A., & Yang, X. (2005). Student Attitudes and Perceptions Regarding Computerized Testing and the Relationship to Performance in Large Scale Assessment Programs. *Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, CA.*

Gu, X. and Rosenbaum, P. R. (1983), "Comparison of multivariate matching methods: structures, distances, and algorithms,"

Ho, D., Imai, K., King, G., and Stuart, E. (2007), Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference, *Political Analysis, 15, 199-236,* http://gking.harvard.edu/files/abs/matchp-abs.shtml.

Ito, K., & Sykes, R. C. (2004). Comparability of Scores from Norm-Referenced Paper-and-Pencil and Web-Based Linear Tests for Grades 4-12. *Paper presented at the annual meeting of the American Educational Research Association,* San Diego, *CA.*

Karkee, T., Kim, D., & Fatica, K. (2010). Comparability Study of Online and Paper and Pencil Tests Using Modified Internally and Externally Matched Criteria. *Paper presented at the annual meeting of the American Educational Research Association,* Denver, CO.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluations of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment, 3(6),* http://www.jtla.org.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment, 2(6),* http://www.jtla.org.

R Development Core Team (2008). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-9000051-07-0, URL, http://www.R-project.org.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-45.

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). Score comparability of online and paper administration of the Texas Assessment of Knowledge and Skills. *Paper presented at the annual meeting of the National Council on Measurement in Education,* San Francisco, CA.

**Table 1**
**Demographic Characteristics of Students**

| *Grade 6* | *Paper-pencil (N=16,830)* | *Online (N=6,551)* | *Total (N=23,381)* |
|---|---|---|---|
| Gender | | | |
| 1.Male | 8609 | 4192 | 12801 |
| 2.Female | 8221 | 2359 | 10580 |
| Race | | | |
| 1. American Indian/Alaskan Native | 86 | 77 | 163 |
| 2. Black | 3512 | 1076 | 4588 |
| 3. Hispanic | 863 | 553 | 1416 |
| 4. Asian | 463 | 182 | 645 |
| 5. White | 11585 | 4532 | 16117 |
| 6. Multiracial | 302 | 130 | 432 |
| 7. Missing | 19 | 1 | 20 |

| *Grade 9* | *Paper-pencil (N=16,969)* | *Online (N=5,605)* | *Total (N=22,574)* |
|---|---|---|---|
| Gender | | | |
| 1.Male | 8629 | 3608 | 12237 |
| 2.Female | 8340 | 1997 | 10337 |
| Race | | | |
| 1. American Indian/Alaskan Native | 107 | 54 | 161 |
| 2. Black | 3445 | 949 | 4394 |
| 3. Hispanic | 773 | 411 | 1184 |
| 4. Asian | 471 | 106 | 577 |
| 5. White | 11879 | 3971 | 15850 |
| 6. Multiracial | 271 | 110 | 381 |
| 7. Missing | 23 | 4 | 27 |

**Table 2**
**Social Studies Blueprint: Grades 6**

| Strand | MC Items | Total Items | Field Test Items | Total Points |
|---|---|---|---|---|
| History | 19 | 19 | * | 19 |
| Geography | 7 | 7 | * | 7 |
| Civics | 10 | 10 | * | 10 |
| Economics | 7 | 7 | * | 7 |
| Knowledge, Processes, Skills | 2 | 2 | * | 2 |
| Total | 45 | 45 | 15 | 45 |

**Each field-test form consists of items from the appropriate strands as determined by a test developer*

**Table 3**
**Social Studies Blueprint: Grades 9**

| Strand | MC Items | Total Items | Field Test Items | Total Points |
|---|---|---|---|---|
| History | 23 | 23 | * | 23 |
| Geography | 13 | 13 | * | 13 |
| Civics | 3 | 3 | * | 3 |
| Economics | 5 | 5 | * | 5 |
| Knowledge, Processes, Skills | 0 | 0 | * | 0 |
| Total | 44 | 44 | 15 | 44 |

**Each field-test form consists of items from the appropriate strands as determined by a test developer*

**Table 4**
**Means and SDs of Latent Trait and Scale Scores for Online and Paper-Pencil in Social Studies Grade 6 and 9**

| Mode | N | Grade 6 Latent Trait | | Grade 6 Scale Score | | N | Grade 9 Latent Trait | | Grade 9 Scale Score | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD | | Mean | SD | Mean | SD |
| Online | 6,551 | -.091 | .642 | 621 | 22.51 | 5,605 | -.042 | .534 | 929 | 25.5 |
| Paper | 6,551 | -.083 | .564 | 623 | 21.14 | 5,605 | -.033 | .423 | 934 | 23.5 |