DOCUMENTOS DE TRABAJO

Saber Investigar

Instituto Colombiano para la Evaluación de la Educación

N° 1

On the comparability of scores from paper- and computer-based achievement tests: Challenges and findings from quasi-experiments

Adrián Quintero
Richard Shavelson
Andrés Rodríguez
Ricardo Duplat
Alexander Calderón

Julio de 2022

icfes
Juntos transformamos el saber

DOCUMENTOS DE TRABAJO
Saber
Investigar

La serie de Documentos de Trabajo Saber Investigar del Icfes tiene como propósito hacer divulgación de los resultados de investigaciones sobre evaluación y análisis de la calidad de la educación.

Serie Documentos de Trabajo Saber Investigar
Edición No. 1
Julio de 2022
Edición digital
ISSN: 2954-6583

Instituto Colombiano para la Evaluación de la Educación
Oficina de Gestión de Proyectos de Investigación
Calle 26 N. 69-76, Elemento, Torre II, piso 18**,** Bogotá, D. C.
Teléfono: (601) 4841410
oficinadeinvestigaciones@icfes.gov.co
https://www.icfes.gov.co

Directora General
Mónica Patricia Ospina Londoño

Jefe de Oficina De Gestión de Proyectos de Investigación
Clara Lorena Trujillo Quintero

Autores
Luis Adrián Quintero Sarmiento
Richard Shavelson
Andrés Ricardo Rodríguez Nagles
Ricardo René Duplat Durán
John Alexander Calderón Rodríguez

Revisión, edición y corrección de estilo
Oficina de Gestión de Proyectos de Investigación

Diseño y diagramación
Oficina Asesora de Comunicaciones y Mercadeo

Advertencia:
El contenido de este documento es el resultado de investigaciones y obras protegidas por la legislación nacional e internacional. No se autoriza su reproducción, utilización ni explotación a ningún tercero. Solo se autoriza su uso para fines exclusivamente académicos. Esta información no podrá ser alterada, modificada o enmendada.

DOCUMENTOS DE TRABAJO
Saber
Investigar

La serie de Documentos de Trabajo Saber Investigar del Icfes tiene como propósito hacer divulgación de los resultados de investigaciones sobre evaluación y análisis de la calidad de la educación.

Citar este documento así:
Quintero, A., Shavelson, R., Rodríguez, A., Duplat, R. y Calderón, A. (2022). *On the comparability of scores from paper-and computer-based achievement tests: Challenges and findings from quasi-experiments*. (Documentos de Trabajo Saber Investigar No. 1). Instituto Colombiano para la Evaluación de la Educación (Icfes). https://www.icfes.gov.co/web/guest/saber-investigar

Comparabilidad de pruebas en papel y computador: retos y hallazgos a partir de cuasiexperimentos[1]

Adrián Quintero[2], Richard Shavelson[3], Andrés Rodríguez[4], Ricardo Duplat[5], Alexander Calderón[6]

Resumen

En este artículo se analiza la comparabilidad de aplicar exámenes en formato de papel y lápiz con los resultados obtenidos cuando se administra la prueba en computador. Para esto se utilizan los datos de la prueba piloto de Saber 3°, 5° y 9° aplicada en Colombia en 2019. En el estudio, la prueba electrónica se implementó en las escuelas con disponibilidad de recursos tecnológicos y en las demás escuelas se aplicó el examen en papel. Por lo tanto, hay diferencias importantes entre las dos muestras de estudiantes y se hace necesario implementar una metodología cuasiexperimental. En este artículo se discuten los retos presentes al utilizar métodos cuasiexperimentales en la comparación de formatos (papel versus computador). Se implementaron métodos de pareo previo al análisis de funcionamiento diferencial del ítem (DIF) y se propone un modelo multinivel para estimar los efectos del formato de presentación. En cada paso del análisis se discuten las limitaciones y se implementan estrategias para hacer el mejor uso de los datos con el fin de extraer conclusiones. Se encuentra que hay un decrecimiento del DIF en grados escolares más altos, y los efectos del formato de aplicación varían entre grados, pero son pequeños en general.

Palabras claves: pruebas en computador; pruebas en papel; comparabilidad; efecto de formato

[1] Las ideas, opiniones, tesis y argumentos expresados son de propiedad exclusiva de los autores y no representan el punto de vista del Icfes. Este trabajo es resultado de un proyecto liderado por Adrián Quintero y Alexander Calderón, realizado en colaboración con el profesor emérito Richard Shavelson de la Universidad de Stanford. Nuestro agradecimiento para la Subdirección de Diseño de Instrumentos por acompañarnos en el análisis cualitativo de los ítems en los dos formatos, a Cristian Montaño y Karen Córdoba por sus sugerencias en cuanto a métodos de pareo y a Natalia González Gómez por sus comentarios a la versión final del documento.

[2] Instituto Colombiano para la Evaluación de la Educación - Icfes, aquintero@icfes.gov.co
[3] Universidad de Stanford, richs@stanford.edu
[4] Instituto Colombiano para la Evaluación de la Educación - Icfes, arodriguez@icfes.gov.co
[5] Instituto Colombiano para la Evaluación de la Educación - Icfes, jcalderon@icfes.gov.co
[6] Instituto Colombiano para la Evaluación de la Educación - Icfes, rduplat@icfes.gov.co

On the comparability of scores from paper- and computer-based achievement tests: Challenges and findings from quasi-experiments

Adrián Quintero, Richard Shavelson, Andrés Rodríguez, Ricardo Duplat, Alexander Calderón

Abstract

We assess the comparability of scores from computer- and paper-based testing using data from the Colombian SABER achievement test administered in grades 3, 5 and 9. As with many countries, the schools and their students that have technological facilities for computer testing differed from the schools/students without such resources, where the paper version had to be administered. Consequently, substantial differences are present in the two student samples. In this paper we discuss the challenges posed by this type of study, since it is well known that employing controlled experiments is the best alternative, but in many cases quasi-experiments are the only available option. Therefore, we implement matching methods prior to differential item functioning (DIF) analysis and propose a multilevel model to estimate the format effects by removing the observable differences between matched samples. In each step we discuss the limitations and implement strategies to make the best possible use of the data to draw conclusions. We found a decrease in DIF for higher grades and mostly small format effects that varied across grades between computer and paper.

Keywords: computer-based test; paper-and-pencil test; comparability; format effects

Contents

1 Introduction

Computer-based testing has become widespread because it offers advantages such as test security, administration efficiency, cost reduction (Bennett et al., 2008; Wang et al., 2008), and interactive item modalities to test complex reasoning (DeBoer et al., 2014). Furthermore, the computer-administered version of a test allows for rapid scoring and permits students to take the test asynchronously at schools or in computing centers. As a result, international assessments such as PISA, TIMSS, PIRLS, ICCS, and national assessments are moving to computers in many countries. However, computer-based tests (CBT) also involve challenges (e.g., Noyes & Garland, 2008) such as access to computers and the internet, similarity in testing conditions (computer resolution, screen size, font size, etc.) and ensuring comparability with paper-and-pencil-tests (PPT) when both administration formats are used simultaneously.

The Ministry of Education in Colombia, for example, is encouraging a transition from PPT to CBT in all national tests. As in many countries, most Colombian schools do not have the necessary resources and connectivity to move to computer-based assessment. Therefore, students in some schools would be administered the test on computers and others with paper-and-pencil. For example, few rural schools have the connectivity necessary to administer CBT, whereas most private schools with high socioeconomic status possess the required facilities. The question immediately arises, “Do the two types of delivery modes produce equivalent (exchangeable, comparable) scores for all students and schools?

Saber 3°, 5°, 9° is a test administered periodically to basic education students in grades 3, 5 and 9 in Colombia. The examinees are assessed in critical reading, mathematics, citizenship, and natural science. The test results are used to analyze national performance over time and to compare the achievement of different regions in the country. The Colombian Institute for Educational Evaluation (Icfes) administered Saber 3°, 5°, 9° in 2019 to a sample of students to determine if scores coming from PPT and CBT may be considered as exchangeable. In this sample, CBT was administered only in schools with computer facilities and connectivity to ensure that the students could answer the test on the computer. Therefore, the examinees in CBT were not randomly selected or randomly assigned to delivery format from the population; they corresponded to a convenience sample of students enrolled in schools with technological facilities.

Although the best way of comparing CBT and PPT is clearly based on two random samples from the whole population of students, this study design is generally difficult to carry out either because of unequal access to IT between schools or because of limited access to administration centers due to cost and logistics, especially in large and diverse countries such as Colombia. Therefore, in this paper we aim to implement a variety of techniques to enable us to analyze the comparability of the two formats and determine whether it is possible to draw conclusions from convenience samples. Based on these analyses, we recommend methods for selecting and analyzing data from the samples in this type of study while balancing practical cost and logistics of administration and a sound way for drawing relevant conclusions.

1.1 Studies comparing CBT and PPT

According to Berman et al. (2020) comparability implies that, ideally, students with the same score are equally proficient in the knowledge which the test intends to measure, regardless the type of administration. In other words, a student would receive the same score if the test were administered in CBT or PPT. A number of studies have investigated the equivalence of the two formats. Some have found that PPT and CBT scores are comparable (Bridgeman et al., 2003; Poggio et al., 2005), while other studies have concluded that scores from the two administration modes are not equivalent (Carlbring et al., 2007; McCoy et al., 2004; Pommerich, 2004). Choi and Tinkler (2002) concluded that items on CBT were more difficult compared to the same items on PPT. The differences were larger for third graders than for tenth graders, and more pronounced in reading than in mathematics. Bennet et al. (2008) found that the mean scale score for eight-graders was significantly lower in CBT compared to PPT, but the difference is very small in effect-size terms. Way et al. (2008) summarized multiple K-12 studies and concluded that several trends tend to emerge depending on the subject being assessed and the grade of the students, but with slight or no format effects in most cases. Piaw (2011) compared CBT and PPT for two psychological tests and concluded that CBT mode was more interpretable as indicated by internal and external validities, besides reducing testing time and increasing examinees' motivation. Nevertheless, students in CBT did not achieve higher scores, on average, than those in PPT. Finally, Jerrim (2016) analyzed PISA 2012 scores in 32 countries/economies and found lower average scores for CBT compared to PPT in 11 of them and higher scores in CBT for 13, but the magnitude of the differences was modest in most cases.

It is also important to analyze equivalence of the scores from CBT and PPT across subpopulations. If the lack of comparability is associated more with individual differences, the change of format could unfairly impact the scores for some groups more than for others. In general, the findings are complex as delivery mode interacts with gender, ethnicity, social class, and access to computers. Some studies have found no important differences in paper vs. electronic delivery performance by gender and ethnicity (Bennett et al., 2008; Way et al., 2008), while others have found significant differences by gender (Jeong, 2014; Jerrim, 2016). There appears to be a trend toward greater comparability with increasing grade level, such that differences by delivery mode are smaller in high school than elementary school, where students obtain higher scores in PPT (Choi & Tinkler, 2002; Hardcastle et al., 2017). Additionally, factors such as socio-economic status may be related to computer familiarity and impact the results in CBT (Bennett et al., 2008).

1.2 Quasi-experimental studies comparing CBT and PPT

It is well known that the best way of comparing CBT and PPT is based on controlled experiments (Shavelson & Towne, 2001), but as explained above, this study design is especially difficult to carry out with school students due to access to IT. For this reason, a number of studies have implemented quasi-experimental designs to compare the two administrations formats. This is the case of Way et al. (2008), Seo and De Jong (2015), Hardcastle et al. (2017) and the references therein. A paramount issue in this type of study is selection bias: at the outset, students taking CBT are not entirely similar to those taking PPT due to access to IT as they tend to attend better financed schools, belong to higher social class and score higher on achievement tests. Simply put, the observed CBT vs. PPT differences may be due to pre-existing differences in students and not delivery mode.

In general, the standard alternative for selection bias in quasi-experimental studies is to employ matching techniques to create equivalent groups in CBT and PPT using students' previous achievement scores and demographic information (Seo & De Jong, 2015; Way et al., 2008). Hardcastle et al. (2017), for example, used propensity-score matching to identify comparable student groups in both testing modes. They used gender, ethnicity, region of the country, and whether English was the student's primary language "… as covariates to calculate a propensity score for each student in each group, and multi-group matching was used to form equivalent groups" (Hardcastle et al., 2017, p. 5). This study reflects the challenge posed by selection bias: a total of 33,422 students in two different CBT groups and one PPT group (three groups in all) provided usable scores in the study. If the three groups were roughly equivalent at the outset, we might expect about 11,140 students in each matched group. However, after propensity score matching, Hardcastle et al. (2017) were able to match 4,959 students in each group. Consequently, the inference from Hardcastle's study should be to populations reflected by these selected students, and not to the full populations reflected in the different groups.

1.3 Present study

The aims of this study are to determine whether it is possible to administer PPT and CBT formats in different schools with different characteristics and: (1) obtain exchangeable scores that allow for comparability at the student level in Saber 3°, 5°, 9° and (2) determine the magnitude of unmatched students which threatens the external validity of findings. We do so in a series of four steps: (1) we employed a propensity-score procedure, genetic matching (Diamond & Sekhon, 2013), to reduce the selection bias present in the two samples. With the matched samples, (2) we then turned to the comparison of the two formats. Our initial comparison of formats was done item-by-item using differential item functioning (DIF). We then examine, qualitatively, items showing large DIF. Then, (3) we compute student scores via item response theory and fit a multilevel model to assess comparability of CBT and PPT scores. And finally, (4) we examine at each step the loss of students in CBT and PPT who could not be matched.

This study, then, addresses five overarching questions:

(1) Does propensity score matching produce equivalent groups and subgroups of CBT and PPT students, and if so, how do the retained students differ from the other students removed from the comparison?

(2) If the samples obtained from matching are not equivalent, what methods can be used to compare the two formats? What is their impact on number of students retained and dropped from the samples?

(3) How many test items showed differential functioning for PPT and CBT? Is there any trend across grades and subjects? Are there consistent item characteristics that give rise to DIF?

(4) For the matched groups, are CBT and PPT scores exchangeable at the student level?

(5) Overall, what are the generalizability of our findings due to loss of student data in the various steps?

When answering these questions, we take into account the limitations caused by having a quasi-experimental design and implement strategies that help obtaining conclusions based on the available data.

2 Materials and methods

2.1 Saber test

The test Saber 3°, 5°, 9° is designed for students enrolled in elementary and secondary schools in Colombia in grades third, fifth and ninth. In November of 2019, this test was administered using between 10 and 15 forms in PPT (depending on the grade and subject) and one of those forms was administered in CBT. The students in third grade responded to two tests, language and mathematics, while students in fifth and ninth grades were administered all the four subjects, namely language, mathematics, citizenship and natural science. For fifth and ninth grades, any single student received tests in 3 of the 4 subjects. The design of this study offers the opportunity to investigate comparability between CBT and PPT across grades at elementary and secondary levels.

Third and fifth grade students received 30 items on each test and ninth-graders 36. All items were multiple-choice, selected-response questions. Together with the Saber achievement test, the students answered a background questionnaire to gather information about socio-economic conditions and computer usage at home.

The schools were contacted via email and telephone two months before the test administration to ask about the number of students, classrooms, computer rooms and the total number of available computers. All administrations, whether PPT or CBT, were proctored by a logistics company hired by Icfes. Students in third grade had two hours and a half to respond the test, while students in fifth grade had three hours and twenty minutes and ninth graders had three hours and fifty minutes.

The computer version of the test was administered using the PLEXI platform designed by Icfes, where the items looked as close as possible to the design in PPT. For CBT, the students had to click the right answer, whereas for PPT the students filled the circle in a paper sheet with the possible answers for the whole test. After finishing the test, the answers from CBT were immediately stored in the cloud. In the paper version, the answer sheets went to a scantron and the data set with the responses was later sent to Icfes for scoring.

2.2 Participants

The students participating in this study were selected using a representative probabilistic multi-stage sampling design. In the first stage, a sample of 301 schools was drawn using a stratified sampling design. Schools eligible to be selected were those with at least five students in one of the three assessed grades. From these 301 schools, only 97 had resources and connectivity for computer-based assessment. The form which was common for PPT and CBT was administered on computers in those 97 schools and administered in paper in another 197 schools. In total, 12,320 students were assessed in the three grades in these 294 schools using the common form. An option to have a randomized trial could be to administer CBT and PPT within the schools with resources and connectivity for computer-based assessment. However, this would imply that the conclusions could be drawn only for students in that type of schools and an important part of the student population would be left out.

Table 1 presents the number of students per administration mode in each grade. The sample size is similar for CBT and PPT in fifth and ninth grades and much larger for the paper version than for the computer assessment in third grade. There were in total 5,659 students assessed in CBT and 6,661 in PPT. The full data set is available from the authors upon request.

Table 1. Number of students in each format per grade.

Grade	CBT	PPT
Third	1,257	2,024
Fifth	2,140	2,508
Ninth	2,262	2,129
Total	5,659	6,661

From the background questionnaire, we characterized the students in CBT and PPT according to their gender, age and some variables related to computer familiarity such as having a computer at home and the number of hours that the student dedicates to using the internet daily for non-academic activities. A socioeconomic index (SEI) was computed for the students based on several questions related to parents' education level, home possessions, type of food consumed generally, the number of persons living at home and number of available rooms. The percentage of missing observations for each variable is between 0.4% and 7.1%. These missing observations were imputed based on fully conditional specification, where each column is imputed, given the values of the other columns in the data. For this, we used the R package MICE (Buuren & Groothuis-Oudshoorn, 2011).

At the school level, there are differences in the performance of private and public schools in Colombia, and whether they are in urban or rural areas. These variables are important in describing the samples. On the other hand, scores for Saber 3°, 5°, 9° are available for 2016 in language and mathematics. We computed the average score for each school in 2016 as a predictor for students' mean achievement in 2019. The average is the mean of the scores for the two subjects and three grades in each school.

Tables A.1-A.3 in the Appendix present a comparison of the CBT and PPT samples for the three grades. For categorical variables, we report percentages; for continuous covariates, we present averages. Together with the difference between the two groups we report if such difference is statistically significant at a 5% significance level (*) and the effect size before matching (ES). As we can see, there are significant differences between students in the two administration modes for some variables, especially for third grade. However, most effect size differences are relatively small, using the criterion of 0.2 as a small effect (Cohen, 1988). In third grade, at the student level we find medium and large differences in age, access to internet at home, computer possession, and SEI. In fifth grade, at the student level, there are medium differences for number of hours of internet use, whether the student has a computer/or internet at home, and SEI. In ninth grade, all variables have a small effect size at the student level. At the school level, for third grade we see difference in: type (public or private), location (rural or urban), and achievement in 2016, whereas there is only one variable with a medium effect size for fifth and ninth grades, respectively, type and school achievement for 2016. This reflects the contrast between students/schools with and without available technological facilities for computer assessment, especially for the lower grades, since for ninth grade there was only one variable with a medium effect size (school achievement in 2016).

2.3 Genetic matching

Students assessed in CBT and PPT differed, especially in third and fifth grades. Therefore, we begin by implementing propensity-score matching, specifically a genetic matching method to obtain samples as comparable as possible. So here, we address the first research question:

(1) Does genetic matching produce equivalent groups and subgroups of CBT and PPT students, and if so, how do the retained students differ from the other students removed from the comparison?

Many different approaches have been used in an attempt to create equivalent groups over the past 20 years (Diamond & Sekhon, 2013). The authors reviewed various approaches and developed an algorithm that integrated them: “genetic matching (GenMatch), eliminates the need to manually and iteratively check the propensity score. GenMatch uses a search algorithm to iteratively check and improve covariate balance, and it is a generalization of propensity score and Mahalanobis distance (MD) matching” (p. 932).

More specifically, after carrying out a matching procedure, it is important to verify that the treatment and control groups have similar joint distributions for the observed covariates. This implies that the distribution for each confounder is close in the two groups. The quality of the matching can be assessed using descriptive statistics for the covariates such as means, maximum and minimum scores, among others. Diamond and Sekhon (2013) proposed a genetic matching method to iteratively check and improve covariate balance using an evolutionary algorithm. Propensity score matching (Rosenbaum & Rubin, 1983) and matching by Mahalanobis distance (Rosenbaum & Rubin, 1985) are particular cases of this procedure. Genetic matching assigns a different weight to each covariate in order to find the particular metric that maximizes post-matching covariate balance. The algorithm minimizes the differences between individuals in the treatment and control groups using a generalized version of the Mahalanobis distance, which includes an additional weight matrix indicating the relative importance of each covariate[7].

While it is desirable to have as many variables as possible for matching, especially individual students’ prior achievement as we are comparing achievement scores from CBT and PPT, we had the variables in Tables A.1-A.3 available for the matching at either the individual or school level. Therefore, we implemented genetic matching for each grade and subject using those variables. Besides genetic matching, we considered nearest neighbor matching based on the logistic regression propensity score, but this led to removing less than 10% of the students in the original samples. Consequently, the differences observed in Table 2 were almost the same after matching with this method. On the other hand, genetic matching

[7] CBT was selected as the treatment group for matching and PPT as the control group. Taking PPT as the treatment group led removing fewer students, but the differences in the covariates between the matched groups remained almost the same.

removed more students but also helped reduce the differences present in CBT and PPT samples.

2.4 Analysis with unbalanced samples

After implementing different matching methods, the resulting samples are still not well balanced for grades third and fifth given the strong differences between students in CBT and PPT. Therefore, it is not possible to carry out a direct quasi-experimental analysis. This leads us to the second research question:

(2) If the samples obtained from matching are not equivalent, what methods can be used to compare the two formats?

To address this question, we consider two approaches to compare CBT and PPT. First, we carry out an item-by-item analysis to study possible differential item functioning (DIF), and then, we scale the two format tests and fit a multilevel model to determine the effect of CBT on student achievement scores. In this multilevel model, we include the covariates that present differences in CBT and PPT after matching to further reduce selection bias and provide a “fairer” comparison of the format effect by removing the unbalancing effects in the data.

DIF analysis

We carried out an item-by-item analysis to study if items behave differently in CBT and PPT. In the ideal case, there are no differences at the item level, so test results for students are equivalent in the two formats. Here, we expect few items to present differences in CBT and PPT because an effort was made to make the items look essentially the same on paper and on screen.

We conducted DIF analysis using the noncompensatory differential item functioning (NCDIF) index proposed by Raju et al. (1995). This method can be employed to study DIF between populations with different abilities since, as explained below, the method equates the item characteristic curves for the two groups under comparison. Therefore, we believe that this method for DIF can give insightful findings in this study for CBT and PPT, even though some differences remain between the samples after genetic matching in third and fifth grades. When implementing NCDIF, the groups under study are matched on students' total scores, so this allows the two groups to have different average abilities, as it is the case here. Therefore, we have two matching procedures involved in this DIF analysis: genetic matching which is based on the covariates, and a posterior matching procedure based on the student abilities in the two groups for NCDIF. It is crucial to complement the DIF analysis here with qualitative methods that critically examine the results obtained.

NCDIF is computed as the integral of the difference between the item characteristic curve (ICC) of the two groups. We present the definition of this statistic for a 2PL model because we use such model to scale the two formats later in this paper. The NCDIF index for each item i is then defined as

$$NCDIF_i = \int_{-\infty}^{\infty} \left(P\left(U_{ij} = 1 \middle| \theta, a_{iF}, b_{iF}\right) - P\left(U_{ij} = 1 \middle| \theta, a_{iR}, b_{iR}\right)\right)^2 f_F(\theta) d\theta$$

where $P\left(U_{ij} = 1 \middle| \theta, a_{iF}, b_{iF}\right)$ is the probability of correctly answering item i for students in the focal group and $P\left(U_{ij} = 1 \middle| \theta, a_{iR}, b_{iR}\right)$ is the probability of correctly answering for the reference group. The integral corresponds to the squared difference of the two ICC's and it is computed over the focal group distribution. The parameters a_{iF}, b_{iF} and a_{iR}, b_{iR} are the discrimination and difficulty parameters for the item according to the 2PL model in the focal and reference groups, respectively. Here, we consider CBT as the reference group and PPT as the focal group.

In order to compute NCDIF, the parameters of the 2PL model are estimated independently for the focal and reference groups. Then, the ICC's are equated so they are in the same scale and the integral of the squared difference for the two groups is computed. We use Stocking and Lord's (1983) method to estimate the equating coefficients. After computing NCDIF for each item, the value can be classified as negligible, moderate, and large DIF according to an effect size measure as proposed by Wright and Oshima (2015). Then we address our third research question:

(3) How many test items showed differential functioning for CBT and PPT? Is there any trend across grades and subjects? Are there consistent item characteristics that give rise to DIF?

Multilevel modeling for the impact of test format

The students item responses were scored using a 2PL Item Response Theory (IRT) model in the R package Mirt (Chalmers, 2012). For this, all items were initially calibrated using the students in CBT, since, as reported below, this is the larger group after matching. The students in PPT were linked through the items that were not flagged with DIF by administration format. This ensures that the scores obtained for both formats are in the same scale. The items flagged with DIF were calibrated freely for the PPT examinees, since the DIF analysis indicates that the parameters for these items are probably different in the two groups.

To determine if there exists a significant difference between the scores in PPT and CBT controlling for the covariates that remained unbalanced after matching, we fitted a multilevel model that includes those covariates and a dummy variable which indicates whether the student was assessed in CBT. To define what covariates are not balanced in the matched groups, we recall that an effect size equal to 0.2 is said to be small (Cohen, 1988). To be more conservative, we consider that covariates with an effect size larger than 0.1 after genetic matching should be included in this multilevel model.

The fitted model considers, as the response variable, the score for the students matched in the study and includes a random intercept for the schools as follows

$$y_{ij} = \beta_{0i} + \beta_1 CBT_{ij} + \vec{\beta}_u \vec{X}_{uij} + \varepsilon_{ij}, \qquad (1)$$

where y_{ij} is the IRT score for student j in school i, β_{0i} the random intercept across schools, CBT_{ij} is a dummy variable which is equal to one if the student was administered the test in CBT and zero otherwise, $\vec{X}_{uij}$ is a vector with the unbalanced covariates and ε_{ij} is the random error in the model. Notice that if $\vec{X}_{uij}$ is not included in the model, the estimate for

β_1 indicates the average score difference between CBT and PPT students. Including those covariates helps finding the average score difference given the covariates which are not balanced in the two format samples. With this methodology we address our last research question:

(4) For the matched groups, are CBT and PPT scores exchangeable at the student level?

3 Results

3.1 Genetic matching

Table 2 shows the sample sizes before and after implementing genetic matching. There was a strong reduction in the sample sizes for PPT, given that CBT was specified as the treatment group and the method removes students in the control group (PPT) to match as well as possible the two groups. The procedure may remove some students in the treatment group (CBT) when there are very atypical observations. The percentage of retained students in PPT after matching is 41% when considering all grades and subjects together. This is a sizable reduction and limits generalizability to the scores that would have been obtained by PPT students dropped from the comparison.

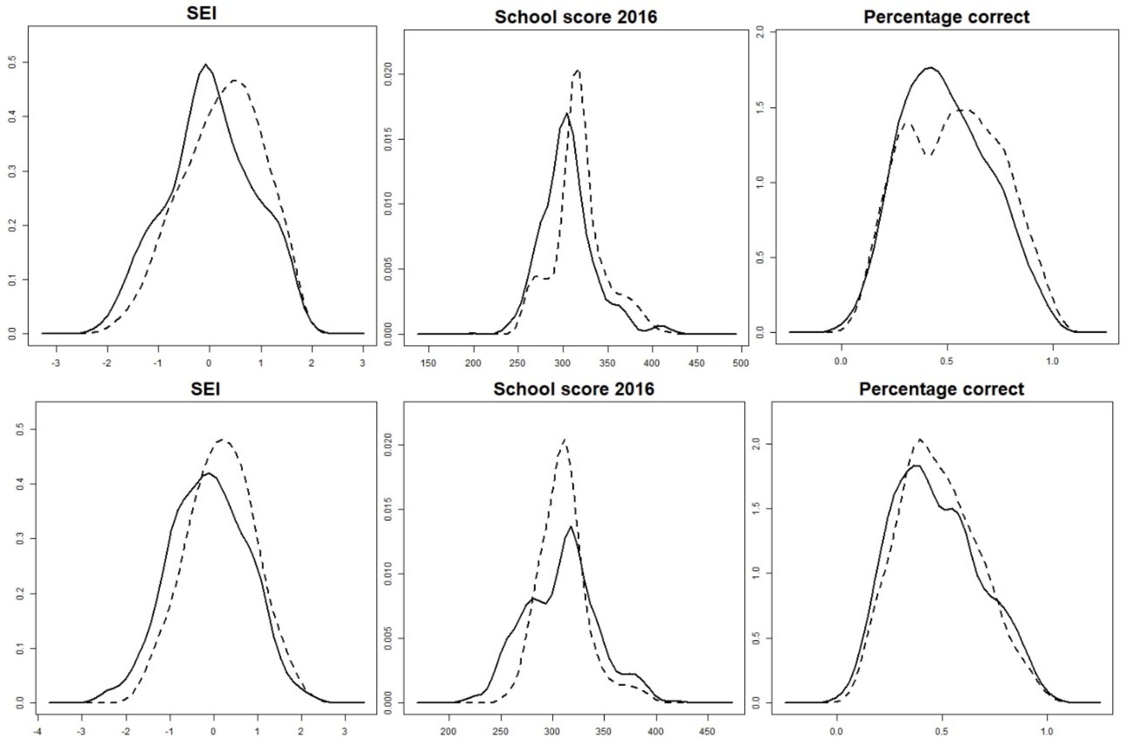
Table 2. Number of students before and after matching.

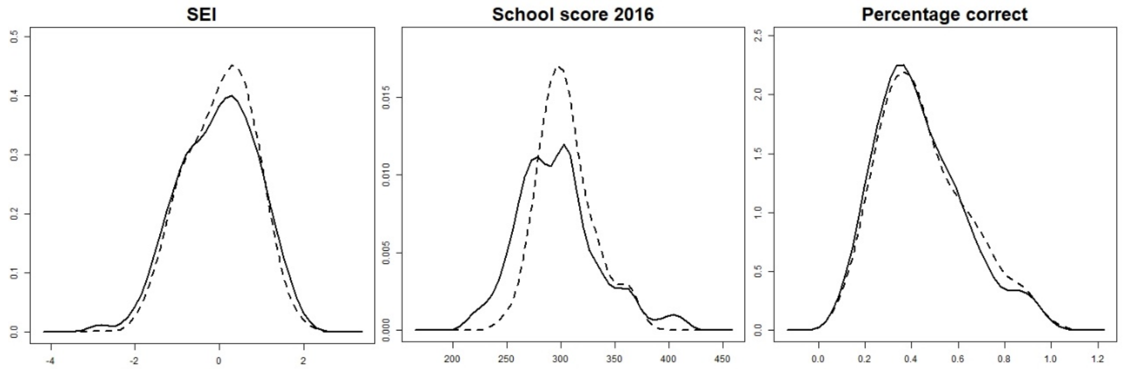
	CBT		**PPT**	
Subject	**Initial**	**Genetic**	**Initial**	**Genetic**
Third				
Language	1,128	1,120	2,000	463
Math	1,246	1,236	1,965	542
Fifth				
Language	1,052	1,052	1,242	545
Math	1,115	1,115	1,236	575

Civics	993	991	1,263	551
Natural S.	1,019	1,017	1,266	568
Ninth				
Language	1,116	1,116	1,069	536
Math	1,143	1,143	1,068	567
Civics	1,098	1,096	1,060	546
Natural S.	1,110	1,107	1,060	524

Figure 1 shows a comparison of retained PPT students in the matching and the other students removed in terms of School score in 2016 and SEI (the two numerical covariates in the study). We present these plots for mathematics, but they were similar across subjects. The full results for the other subjects are available from the authors upon request.

Figure 1. Density plots for SEI, school score in 2016 and percentage of correct answers comparing for matching.





Note: Retained students (dashed line) and discarded students (solid line) for math. On the top third graders. in the middle fifth graders and ninth graders on the bottom.

There are clear differences between the two groups, since the retained students tend to belong to schools with higher achievement in 2016 and with higher SEI (Tables A.1-A.3). This is as expected, since students in CBT have these characteristics, and the matching selects similar students in PPT. We also present the distribution of the percentage of correct answers for the retained and removed students (Figure 1). The retained students tend to have a higher performance, on average.

These results occur often in this type of quasi-experiments for comparing CBT and PPT, since the schools and their students that have technological facilities for computer testing strongly differ from the schools/students without such resources. For instance, the percentage of retained students in Hardcastle et al. (2017) was around 43%, which is very similar to the value in this study (41%). This is problematic, since a large part of the sample is being discarded in a non-random manner and the consequences of that are not clear. However, something positive is that, as observed in Figure 1, the distribution of retained students' performance covers the same interval as removed students. Consequently, the inferences made based on matched students may apply for low, medium, and high achieving examinees.

Table 3 reports the number of schools in the study before and after implementing genetic matching. For CBT, all schools were retained after matching for all grades and subjects. On the other hand, the number of schools is reduced for PPT, with about 40% of schools removed in third grade, 25% removed in fifth grade and 20% for ninth grade.

Table 3. Number of schools before and after matching.

	CBT		PPT	
Subject	**Initial**	**Genetic**	**Initial**	**Genetic**
Third				
Language	29	29	272	159
Math	31	31	271	174
Fifth				
Language	46	46	262	194
Math	49	49	262	195
Civics	46	46	262	208
Natural S.	46	46	262	203
Ninth				
Language	49	49	256	187
Math	51	51	256	205
Civics	50	50	258	200
Natural S.	50	50	258	195

Tables A.1-A.3 in the Appendix report the effect size for the covariates before matching (ES) and after matching (ES-match) in the three grades. A matching procedure was carried out independently for each subject since the students were administered tests in different subjects. We present the effect size after matching in Tables A.1-A.3 in the Appendix for mathematics, but the results are very similar across subjects. Effect size differences were reduced after matching as expected. In fifth and ninth grades, all effect size differences after matching are small (values equal or lower than 0.2), except for SEI in fifth grade. As pointed out before, the covariates are very unbalanced for third grade before matching. The matching procedure improved the balancing but the covariates School score in 2016 and School type still present a medium effect size after genetic matching.

Keeping these challenges in mind, we proceed to the following analyses with the matched samples taking care of the possible conclusions that can be made under the present restrictions. For DIF analysis we use NCDIF, which implements an additional matching procedure based on the student scores for the two groups and the results are verified using qualitative methods. In addition, to estimate the format effects we include the unbalanced covariates in the model to reduce the impact of such effects in the two samples. As discussed previously, we consider covariates with an effect size larger than 0.1 as unbalanced.

We present the variables with an effect size larger than 0.1 for each grade in Table 4 for the matched students in mathematics (the covariates are practically the same across subjects). As aforementioned, third graders in CBT and PPT present stronger differences, and as so, they present more unbalanced covariates, while ninth graders have fewer unbalanced covariates in the three grades. This is related to school dropout since students with lower socioeconomic conditions are more likely to drop school. As a result, stronger differences are observed between students in lower grades.

Table 4. Covariates with an effect size larger than 0.1 for mathematics after genetic matching.

Grade	Variables
Third	School type, age, Internet at home, Computer at home, SEI, School score 2016
Fifth	School location, School type, Internet at home, Computer at home, SEI
Ninth	School location, School type, School score 2016

3.2 DIF analysis

Table 5 reports the DIF magnitude according to the effect size measure of NCDIF. As expected, few items present large DIF (C), since an effort was made when designing the two format tests to make the items look essentially the same on paper and on screen. The number of items with moderate DIF (B) is also not large. However, there is a pattern in which the number of DIF items decreases with grade level.

Table 5. Number of items with negligible (A), moderate (B), and large (B) DIF magnitude according to the effect size measure of NCDIF

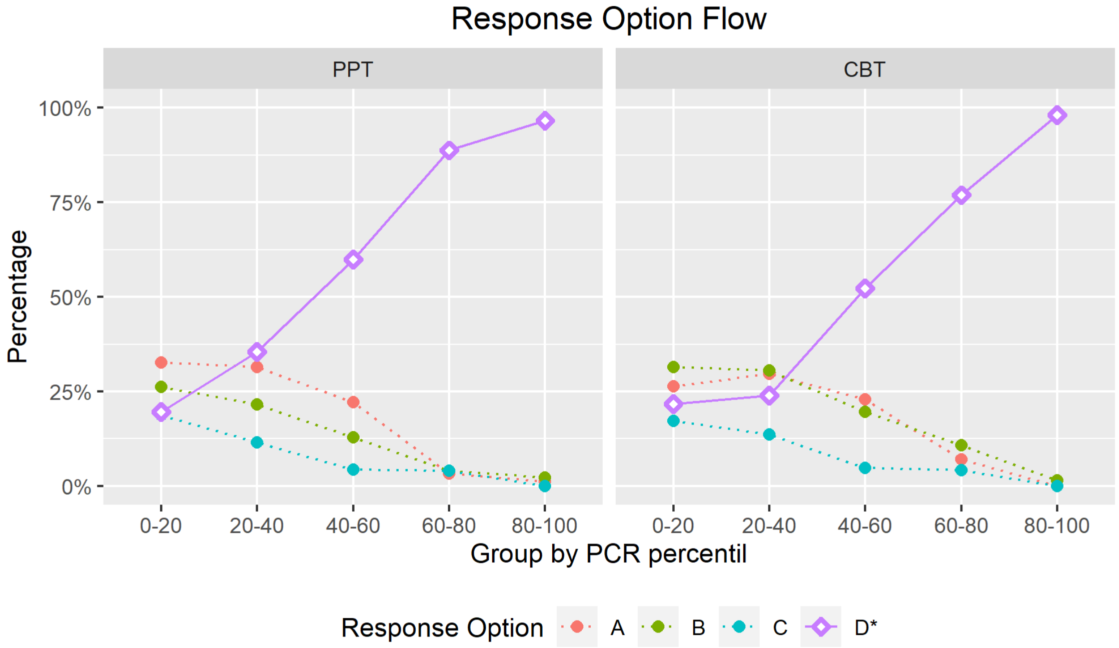
Subject	Total items	A	B	C
Third				
Language	26	16	7	3
Math	24	12	7	5
Fifth				
Language	29	25	3	1
Math	25	21	3	1
Civics	23	20	2	1
Natural S.	28	23	3	2
Ninth				
Language	29	26	2	1
Math	22	20	2	0
Civics	31	29	2	0
Natural S.	31	30	0	1

If the number of items with large DIF was considerable, it would be difficult to argue that the total score is a meaningful matching criterion between the two groups in NCDIF since the scores from the two formats would probably have a different interpretation. However, this is not the case here and we can assume that the two formats correspond to the same test.

On the other hand, the matched samples for CBT and PPT still have some differences as discussed above. We believe that the impact of these differences is not substantial since NCDIF matches the two groups based on the students' scores, such as it is done in other cases when DIF is analyzed for gender, for instance. In that case, one of the two groups may have higher ability and the characteristic curves of the two groups are equated before drawing conclusions.

All items were carefully reviewed with expert test designers of each subject to explore the results obtained in this DIF analysis based on how they look in CBT and PPT. A useful tool to better understand why DIF was observed in the items is the flow response options plot (Figure 2), which clusters the students in the x-axis according to the percentage of correct responses (as a proxy for their ability), and, in the y-axis, it reports the percentage of students in each cluster that selected each option (A, B, C, D) when answering the item. The display helps us understand what options of the item are behaving different in the two formats and for what ability levels.

Figure 2. Flow response options plot for an item of mathematics third grade.



Note: The key is the option B with an asterisk.

The conclusions about the reasons for DIF varied depending on the subject but, in general, the differences were mainly observed for low and medium-low ability students. For instance, in Figure 2, for the lowest ability group (0-20), the students selected each option with probability close to 25%, and the probability of selecting the key remained almost the same for the following group (20-40) in CBT, while for PPT such probability increased to 37%. In general, there was not a clear trend since about half of the DIF items favored PPT and the other half favored CBT.

When analyzing the items in detail and comparing them between formats, it was found that they look very similar in CBT and PPT, so it is difficult to understand the causes of DIF. However, for language and natural science, most items with DIF required scrolling for reading/selecting the option responses and a clear trend emerged. When the key was A or B, the students in CBT tended to have a higher percentage of correct responses, since these are the first options the student sees when scrolling, and sometimes they do not scroll until they see all four response options. On the other hand, when the key was C or D, the students in CBT tended to have a lower percentage of correct responses (as in Figure 2). This confirms previous findings about scrolling (Choi & Tinkler, 2002; Way et al., 2008),

For mathematics and civics the items did not require scrolling in general, and it was more difficult to understand the causes of DIF. However, when DIF was observed, the flow response options plot suggested that it was mainly observed in low and medium-low ability students. It could be caused because those students have more issues or get more easily distracted by some option responses when answering an item in one format or the other. An additional study is necessary to understand well this phenomenon.

3.3 Multilevel modeling for the impact of test format

In the scoring process, all items were initially calibrated with a 2PL model using the students in CBT. Then, the students in PPT were linked through the items with a negligible DIF effect. Table 6 reports the raw mean scores for CBT and PPT without taking into account the unbalanced covariates for the two samples. With increasing grade level, CBT versus PPT mean scores switch from favoring CBT to PPT. The effect size is small in all cases, except for mathematics and language in third grade and civics in ninth grade, where the effect size is medium. The variability of scores is similar across grades except for smaller variability in CBT scores for third grade language, fifth grade language and natural science, and ninth grade civic competences.

Table 6. Average and standard deviation of IRT scores for matched students in CBT and PPT

Subject	Mean CBT	Mean PPT	Mean difference	Cohen's d	SD CBT	SD PPT	SD ratio
Third							
Language	50.00	47.31	2.69*	0.26	9.04	9.60	0.94*
Math	50.00	46.05	3.94*	0.40	8.89	9.11	0.98
Fifth							
Language	50.00	48.67	1.33*	0.14	9.13	9.74	0.94*
Math	50.00	49.04	0.96*	0.11	8.97	8.92	1.01
Civics	50.00	49.54	0.46	0.05	8.59	9.03	0.95
Natural S.	50.00	49.02	0.98*	0.10	9.11	9.84	0.93*
Ninth							
Language	50.00	50.86	-0.86*	-0.10	8.97	9.10	0.99
Math	50.00	51.56	-1.56*	-0.18	8.54	8.92	0.96
Civics	50.00	52.04	-2.04*	-0.22	8.97	9.87	0.91*
Natural S.	50.00	50.49	-0.49	-0.06	9.10	8.93	1.02

Note: *Statistically significant at a 5% level

Generally, the average difference between the two formats (e.g., 2.69 for language in third grade) is considered as the correction factor to be taken if we intend to have comparable scores between the two formats (e.g., Way et al., 2008). However, those differences reflect not only the format effect but also the covariates which are not yet well balanced in the matched samples, and it is important to take them into account as shown below.

When unbalanced covariates are controlled in the comparison using model (1), we find no differences between PPT and CBT in mathematics and language for third graders (Table 7). Similarly, fifth graders have similar scores in the two formats for the four subjects, whereas ninth graders score lower in CBT in all subjects.

Table 7. Impact of format on unbalanced covariate adjusted scores and its effect size.

Subject	CBT effect	Cohen's d
Third		
Language	-0,79	-0,08
Math	0,80	0,08
Fifth		
Language	0,15	0,02
Math	-0,05	-0,01
Civics	-0,09	-0,01
Natural S.	0,17	0,02
Ninth		
Language	-1.37*	-0,15
Math	-1.83*	-0,22
Civics	-2.31*	-0,25
Natural S.	-0.97*	-0,11

Note: *Statistically significant at a 5% level

The use of the multilevel model had its intended result for showing more clearly the format effect after reducing the impact of the unbalanced covariates in the two matched samples. For example, the model suggests no significant differences, on average, between CBT and PPT for third and fifth grades, whereas significant differences were found, on average, for mathematics and language in third grade, and for three of the four subjects assessed in fifth grade (Table 6). The results before adjusting for the unbalanced covariates indicated that CBT was easier than PPT in the lower grades, which is not in agreement with the literature (see e.g., Choi & Tinkler, 2002; Hardcastle et al., 2017).

The samples were not randomly assigned to CBT and PPT in this study, so it is not possible to affirm that the values in Table 7 correspond exactly to the format effect. There could be unobservable variables which are different in the two groups and are also creating differences in the student performance. However, the step-by-step procedure we followed synthesizes the best strategy that we found to make conclusions based on the available data. In addition, as shown in Figure 1, the inferences based on matched students apply for low and high achieving examinees (who were retained in the analysis).

If it was necessary to report student results for CBT and PPT in the same scale, the best alternative in this study would be to use the conversion in Table 7 to adjust the equated scores in the two formats. Such results indicate that ninth graders obtained lower scores in CBT in the four subjects, and the effect size is small for language and natural science, and medium for math and civic competences. The effect sizes are small for third graders and negligible for fifth graders in all subjects.

4 Discussion

The present study collected data from 12,320 students in third, fifth and ninth grades to assess the comparability between CBT and PPT in multiple subjects. The comparison of formats was carried out in a natural quasi-experimental design as neither schools nor students could be randomly allocated to test format given the monetary, logistic and personnel costs that would be encountered. However, the lack of randomization makes the comparison challenging due to systematic differences between the students/schools in the two samples.

To address this selection bias by matching CBT and PPT students, a genetic matching procedure was employed. Even after matching, the samples of students in CBT and PPT still presented some differences. This occurs because the students in schools with and without technological facilities have very different conditions in Colombia, as it is the case in many other countries. Moreover, we experienced a loss of almost 60% of PPT students. Therefore, when implementing DIF analysis, we equated the scores of the remaining students in the two groups to take into account differences in the performances of the two groups. The results showed a decrease in DIF for higher grades. In addition, we included the unbalanced covariates in the multilevel model to estimate the format effects. Such a model showed the format effects for third grade (math and language) as medium to being small and non-significant after removing the effect of the unbalanced covariates.

The estimates in Table 7 indicate that the impact of format is small according to the effect size in all cases, except for math and civics in ninth grade. Given the lack of randomization in the study, it is not possible to ascertain that the estimates in Table 7 reflect only the format effect, but this is the best approximation that we found in this type of quasi-experiment with the available data. The methodology presented here could be implemented in other studies where a quasi-experimental design is the only alternative.

The analyses were accompanied by a detailed qualitative analysis of the items to try to find possible reasons for DIF. The findings confirmed that scrolling definitely has an impact on the comparability of the results and also suggested that, in Colombia, low and medium-low ability students show stronger differences when answering in CBT or PPT even if there are no evident reasons for that when looking at the items.

In order to keep a balance between randomization in the study, generalizability of the conclusions and costs of the experiment, a possibility could be to take the sample within the group of schools that have resources for CBT. However, it is important that the sample of schools covers all ranges of the ability scale without discarding low achieving students, who tend to be enrolled in schools without resources for CBT. The ability of the students in the schools can be approximated from previous achievement tests. If possible, the distribution of the abilities in the selected sample of schools with computers and connectivity should be very close to the distribution of the abilities for all schools at the national level. After selecting such sample of schools, the students could be randomized within each school in CBT and PPT.

The conclusions based on the above design could lead to more solid conclusions compared to a quasi-experimental design. However, there could be some differences between the sample and the overall student population, especially with respect to the familiarity that students may have with computers in schools with and without such resources. Nevertheless, familiarity should not be a factor that creates differences in the performance in CBT as all students should receive previous training for answering the items in computer. Otherwise, the test in CBT might not be only measuring the desired latent variable but also the student's familiarity with computers. As an additional suggestion, it is important to administer a questionnaire that measures student familiarity with computers to analyze the impact that it may have on student scores.

Declaration of interest statement

The authors have declared no conflict of interest.

References

Arikan, S., Vijver, F. J. R. van de, & Yagmur, K. (2018). Propensity Score Matching Helps to Understand Sources of DIF and Mathematics Performance Differences of Indonesian, Turkish, Australian, and Dutch Students in PISA. *International Journal of Research in Education and Science*, 69-82.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). *Does it Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP*. 39.

Berman, A. I., Haertel, E. H., & Pellegrino, J. W. (2020). *Comparability of Large-Scale Educational Assessments: Issues and Recommendations*. Washington, DC: National Academy of Education

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of Screen Size, Screen Resolution, and Display Rate on Computer-Based Test Performance. Applied Measurement in Education, 16(3), 191-205, DOI: 10.1207/S15324818AME1603 _2

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). **mice**: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(1), 1-67.

Carlbring, P., Brunt, S., Bohman, S., Austin, D., Richards, J., Öst, L.-G., & Andersson, G. (2007). Internet vs. Paper and pencil administration of questionnaires commonly used in panic/agoraphobia research. *Computers in Human Behavior*, *23*(3), 1421-1434.

Chalmers, R. P. (2012). **mirt**: A Multidimensional Item Response Theory Package for the *R* Environment. *Journal of Statistical Software*, *48*(6).

Choi, S. W., & Tinkler, T. (2002). *Evaluating comparability of paper-and-pencil and computer-based assessment in a K–12 setting.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), New Orleans, LA. (s. f.). https://nceo.info/references/paper-conference/10706

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

DeBoer, G. E., Quellmalz, E. S., Davenport, J. L., Timms, M. J., Herrmann-Abell, C. F., Buckley, B. C., Jordan, K. A., Huang, C.-W., & Flanagan, J. C. (2014). Comparing three online testing modalities: Using static, active, and interactive online testing modalities to assess middle school students' understanding of fundamental ideas and use of inquiry skills related to ecosystems. *Journal of Research in Science Teaching*, *51*(4), 523-554.

Diamond, A., & Sekhon, J. S. (2013). Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics*, *95*(3), 932-945.

Hardcastle, J., Herrmann-Abell, C. F., & DeBoer, G. E. (2017). *Comparing Student Performance on Paper-and-Pencil and Computer-Based-Tests.* Paper presented at the 2017 AERA Annual Meeting, San Antonio, TX.

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology*, *33*(4), 410-422.

Jerrim, J. (2016). PISA 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice*, *23*(4), 495-518.

McCoy, S., Marks, P. V., Carr, C. L., & Mbarika, V. (2004). Electronic versus paper surveys: Analysis of potential psychometric biases. *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, 8 pp.

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics*, *51*(9), 1352-1375.

Piaw, C. (2011). *Comparisons Between Computer-Based Testing and Paper-Pencil Testing: Testing Effect, Test Scores. Testing Time and Testing Motivation.* In Proceedings of the Informatics Conference at: University of Malaya (pp. 1-9).

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A Comparative Evaluation of Score Results from Computerized and Paper & Pencil Mathematics Testing in a Large Scale State Assessment Program. *The Journal of Technology, Learning and Assessment*, *3*(6), Article 6.

Pommerich, M. (2004). Developing Computerized Versions of Paper-and-Pencil Tests: Mode Effects for Passage-Based Tests. *The Journal of Technology, Learning and Assessment*, *2*(6), Article 6.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician*, *39*(1), 33.

Seo, D. G., & De Jong, G. (2015). Comparability of Online- and Paper-Based Tests in a Statewide Assessment Program: Using Propensity Score Matching. *Journal of Educational Computing Research*, *52*(1), 88-113.

Shudong Wang, Hong Jiao, Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of Computer-Based and Paper-and-Pencil Testing in K–12 Reading Assessments: A Meta-Analysis of Testing Mode Effects. *Educational and Psychological Measurement*, *68*(1), 5-24.

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement*, *27*(4), 361-370.

Towne, L., & Shavelson, R. J. (2002). *Scientific research in education.* Washington, DC: The National Academies Press. https://doi.org/10.17226/10236

Way, W. D., Lin, C. H., & Kong, J. (2008). *Maintaining Score Equivalence as Tests Transition Online: Issues, Approaches and Trends*. 33. Paper presented at the e annual meeting of the National Council on Measurement in Education, New York, NY.

Appendices

Table A.1. Comparison of samples in CBT and PPT for third grade, their difference (DIFF), effect size before matching (ES) and after matching (ES-match) for math.

Categories	CBT	PPT	DIFF	ES	ES-match
Sex					
Female	0.49	0.49	-0.01	-0.01	-0.02
Age					
7 years old or less	0.01	0.03	-0.01*	-0.10	-0.04
8 years old	0.40	0.34	0.06*	0.13	0.05
9 years old	0.46	0.44	0.02	0.04	0.05
10 years old or more	0.12	0.19	-0.07*	-0.19	-0.13
Internet hours					
No use	0.19	0.28	-0.08*	-0.20	-0.09
Less than 1	0.31	0.26	0.05*	0.11	0.09
Between 1 and 3	0.23	0.17	0.06*	0.14	0.08
Between 3 and 5	0.09	0.12	-0.03*	-0.10	-0.09
More than 5	0.18	0.17	0.01	0.02	-0.02
Internet at home					
Yes	0.79	0.68	0.12*	0.27	0.11
Computer at home					
Yes	0.71	0.54	0.18*	0.37	0.16
School type					
Private	0.23	0.04	0.18*	0.58	0.35
School location					
Urban	0.99	0.90	0.09*	0.42	0.08
School score 2016	329.80	308.46	21.34*	0.61	0.34
SEI	0.40	0.08	0.32*	0.42	0.20

Note: * p < .05 for the test of the differences in means between CBT and PPT

Table A.2. Comparison of samples in CBT and PPT for fifth grade, their difference (DIFF), effect size before matching (ES) and after matching (ES-match) for math.

Categories	CBT	PPT	DIFF	ES	ES-match
Sex					
Female	0.47	0.50	-0.03*	-0.06	-0.05
Age					
7 years old or less	0.01	0.02	0.00	-0.04	0.09
8 years old	0.32	0.34	-0.02	-0.05	0.01
9 years old	0.49	0.42	0.07*	0.14	0.02
10 years old or more	0.17	0.22	-0.04*	-0.11	-0.06
Internet hours					
No use	0.16	0.26	-0.10*	-0.24	-0.10
Less than 1	0.30	0.29	0.01	0.02	0.00
Between 1 and 3	0.29	0.25	0.04*	0.08	0.02
Between 3 and 5	0.10	0.08	0.02*	0.06	0.02
More than 5	0.16	0.13	0.03*	0.09	0.06
Internet at home					
Yes	0.80	0.70	0.10*	0.24	0.10
Computer at home					
Yes	0.68	0.57	0.11*	0.22	0.13
School type					
Private	0.11	0.04	0.07*	0.26	0.19
School location					
Urban	0.91	0.91	0.00	0.01	-0.19
School score 2016	313.19	309.54	3.65*	0.16	0.09
SEI	0.33	0.00	0.34*	0.43	0.21

Note: * p < .05 for the test of the differences in means between CBT and PPT

Table A.3. Comparison of samples in CBT and PPT for ninth grade, their difference (DIFF), effect size before matching (ES) and after matching (ES-match) for math.

Categories	CBT	PPT	DIFF	ES	ES-match
Sex					
Female	0.53	0.51	0.02	0.04	0.06
Age					
7 years old or less	0.01	0.01	0.00	-0.05	0.09
8 years old	0.27	0.30	-0.03	-0.06	-0.02
9 years old	0.44	0.41	0.03*	0.07	-0.01
10 years old or more	0.28	0.29	0.00	-0.01	0.02
Internet hours					
No use	0.07	0.09	-0.02	-0.07	-0.01
Less than 1	0.13	0.15	-0.01	-0.04	-0.04
Between 1 and 3	0.31	0.33	-0.03	-0.06	-0.02
Between 3 and 5	0.21	0.19	0.02	0.04	0.03
More than 5	0.28	0.23	0.04*	0.10	0.04
Internet at home					
Yes	0.75	0.73	0.02	0.05	0.02
Computer at home					
Yes	0.67	0.62	0.04*	0.09	0.09
School type					
Private	0.07	0.06	0.01	0.03	0.12
School location					
Urban	0.91	0.92	-0.01	-0.03	-0.16
School score 2016	309.02	301.14	7.88*	0.30	0.16
SEI	0.00	-0.02	0.02	0.02	0.00

Note: * p < .05 for the test of the differences in means between CBT and PPT