AERA Open
January-March 2017, Vol. 3, No. 1, pp. 1–11
DOI: 10.1177/2332858417692997

Measuring Student Learning in Technical Programs: A Case Study From Colombia

Benjamin W. Domingue
David Lang
Stanford University
Martha Cuevas
Instituto Colombiano para la Evaluación de la Educación
Melisa Castellanos
Concordia University
Carolina Lopera
Julián P. Mariño
Adriana Molina
Universidad de los Andes
Richard J. Shavelson
Stanford University

Technical schools are an integral part of the education system, and yet, little is known about student learning at such institutions. We consider whether assessments of student learning can be jointly administered to both university and technical school students. We examine whether differential test functioning may bias inferences regarding the relative performance of students in quantitative reasoning and critical reading. We apply item response theory models that allow for differences in response behavior as a function of school context. Items show small yet consistent differential functioning in favor of university students, especially for the quantitative reasoning test. These differences are shown to affect inferences regarding effect size differences between the university and technical students (effect sizes can fall by 44% in quantitative reasoning and 24% in critical reading). Differential test functioning influences the rank orderings of institutions by up to roughly 5 percentile points on average.

Keywords: *technical schools, community colleges, Colombia, higher education, differential test functioning*

Globally, higher education is a crucial component of many students' paths to professional success. However, in only a few countries is the learning that occurs in higher educational institutions systematically assessed (Zlatkin-Troitschanskaia, Shavelson, & Kuhn, 2015). Moreover, existing assessments tend to focus on universities. "Technical schools"—which we define broadly to include community colleges, junior colleges, and technical schools—are an important part of the higher education system that offer both important vocational training and certification and, for many students, preparation for enrollment in traditional institutions of higher learning. There is limited research available about these institutions, especially about their effect on student learning. Expanding this limited research base is important because technical schools serve large numbers of students, many of whom are underrepresented in traditional higher educational institutions. For example, they are likely to be of lower socioeconomic status as compared to students of traditional universities. An additional concern is that for-profit institutions have become major providers in this space and are frequently of dubious quality (e.g., Deming, Goldin, & Katz, 2011). An understanding of the impact these institutions have on learning is clearly needed.

The development of specialized instruments to measure learning outcomes in technical schools would be costly and time-consuming. A natural alternative, especially given the complementary nature of their missions, is to design instruments to measure the learning of both university and technical school students. Although this solution is promising from a budgetary perspective, caution is required because university and technical students may encounter dramatically different pedagogical environments (e.g., study of the thermodynamic principles that underlie engines for university students versus

focus on maintaining a mechanical engine for technical school students).

This study explores the degree to which institutional context may lead to bias in the measurement of student ability. We examine the properties of an assessment for higher education learning outcomes in both universities and technical schools, focusing specifically on evidence regarding measurement invariance (Millsap, 2012) across the two settings. In particular, we focus on the extent to which differential test functioning (DTF; Chalmers, Counsell, & Flora, 2015) may challenge our interpretation of the difference in mean ability after higher education between students attending each type of institution.

To make this comparison, we use data from Colombia. Colombia offers a unique opportunity for such study because there is a universal assessment of quantitative reasoning (QR) and critical reading (CR) competencies at the end of postsecondary schooling, the SABER PRO. Using data from this assessment, we focus on two main questions. First, does the test measure student ability in comparable fashions across the two institutional contexts (i.e., do we observe DTF)? Second, how might DTF influence our perceptions of mean student performance in each type of institution and the differences in school rankings derived from the assessment? Foreshadowing our findings, we provide evidence that the test demonstrates DTF in favor of university students and that this fact has implications—for both the perceived effect size difference between mean student ability at each type of institution and the types of school orderings—that are relevant to both policy and the broader public. Before describing the empirical evidence upon which we based those statements, we begin by situating this study in the broader context surrounding the measurement of learning outcomes in higher education and offer additional information on the SABER PRO and its relevance in Colombia.

Measuring Student Learning in Higher Education

As students around the globe enroll in ever-larger numbers in higher education, there is an increasing need to understand, at a broad level, student learning in such environments (see discussion in Zlatkin-Troitschanskaia et al., 2015). Numerous nations have witnessed the development of such assessments: the Collegiate Learning Assessment (CLA+) in the United States (Klein, Benjamin, Shavelson, & Bolus, 2007; see also http://cae.org/students/college-student/what-is-cla/), the ENADE in Brazil (Verhine, Dantas, & Soares, 2006), the Graduate Skill Assessment in Australia (Hambur, Rowe, & Luc, 2002), and the KoKoHs in Germany (Zlatkin-Troitschanskaia, Pant, Kuhn, Toepper, & Lautenbach, in press). There has also been the development of an international measure of learning in higher education led by the Organisation for Economic Co-operation and Development (the 17-nation Assessment of Higher Education Learning Outcomes partnership; Tremblay, Lalancette, & Roseveare, 2012). We emphasize that most measures of learning in higher education are typically administered to boutique samples of students (e.g., they are either random samples from a small set of institutions or convenience samples) and nearly never in technical schools.

Before describing specifics of the SABER PRO, we think it important to discuss whether the curricula used in technical schools should result in student learning detectable by such an instrument. The curricula used in technical schools are typically designed so as to foster the types of skills and thinking that are useful in specific vocational settings. However, the learning should not be so specific that it would not generalize to other areas of life (e.g., managing one's personal finances) or even other types of careers. Learning that does not generalize in this manner would be job-specific training of a type that may be inappropriate (indeed, in Colombia, technical programs are legally required to develop generic competencies). Furthermore, the SABER PRO was designed to be used in both universities and technical schools, suggesting that, at least in Colombia, there is a belief that technical schools should lead to generalizable gains in student learning.

The SABER PRO

The SABER PRO is part of a suite of educational measures developed by ICFES, a governmental agency with a mission of educational evaluation, for use in Colombia. This assessment system includes a set of assessments administered yearly in all schools to all students in Grades 3, 5, and 9 (the SABER 3, 5, and 9) as well as the SABER 11, which acts as both an indicator for the quality of secondary education and a college admissions exam in Colombia and is aligned with SABER PRO in order to produce value-added indicators. The SABER PRO is a college exit examination that must be taken by all students completing higher education (a certificate of completion is required for graduation). The SABER PRO assesses both generic and domain-specific skills. On the generic skill side, the SABER PRO measures QR, CR, English, and citizenship skills in a multiple-choice format and written communication in a performance setting. On the domain-specific side, it includes tests such as education and law (Shavelson et al., 2016). Test development utilized principles of evidence-centered design (Mislevy & Haertel, 2006). Crucially, instructors from throughout the Colombian higher education system, including both universities and technical schools, were involved in the construct definition, the test specification, and the item writing process. We focus on the generic skills, specifically, QR and CR, as they are most clearly influenced by the curricula in technical schools (English, for example, is unlikely to be a focal point for many technical school students in Colombia) and relevant to technical students. Furthermore, the written

communication and citizenship skills tests also had technical problems that would have complicated their inclusion in this analysis.

Higher Education in Colombia and the Role of the SABER PRO

We begin by describing the higher education system in Colombia. In 2007, 31.6% of all 17- to 21-year-olds attended higher education institutions. By 2015, this percentage had risen to 49.4%. Undergraduate education consists of programs classified into one of three levels: 1, technical (professional technical programs); 2, technological (technology programs); and 3, professional (university professional programs). In this study, we compare technical and technological programs to professional programs. We colloquially describe these programs as "technical schools" and "universities," but note that some institutions can grant degrees of various programmatic types. Technical and technological programs typically require 2 to 3 years, whereas professional degrees take 4 to 5 years of study. Work toward technical and technological degrees can frequently be transferred as credit toward a professional degree, similar to how U.S. students may transfer credit from a community college to a 4-year university.

At present, there are virtually no stakes for the students who take the SABER PRO. On the other hand, results have weighty implications for universities. In particular, unofficial rankings of institutions based on mean SABER PRO scores within an institution have been computed (Bogoya, 2012) and widely reported ("Las instituciones con las mejores pruebas Saber Pro," 2013; "Los Andes, primera en Pruebas Saber Pro," 2013; "Uniandes, la mejor en Pruebas Saber Pro 2012," 2013). Official rankings are now being produced that also utilize SABER PRO scores (Ministerio de Educación Nacional, n.d.). Moreover, value-added measures are being computed for institutions of higher education (Milla, San Martin, & Van Bellegem, 2016; Shavelson et al., 2016). Results based on value-added analyses are also being made publicly available (Instituto Colombiano para la Evaluación de la Educación, n.d.). Given this focus and the need for a test's uses to be considered in a discussion of its validity (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, Standard 1.1), an examination of the degree to which school effectiveness metrics may be sensitive to DTF (due here to the type of higher educational institution) is timely.

Data

We focus on data from the spring 2012 administration of the SABER PRO. Of the 106,189 students who took the SABER PRO in the spring of 2012, we focus on the subset ($n \approx 97{,}000$; see Table 1) that responded to all items. The majority of students obtained professional credentials (i.e., graduated from universities), but 31,192 students obtained technical credentials. Both the QR and CR tests consisted of 35 multiple-choice items, which were dichotomously scored (correct/incorrect) and had reliability coefficient estimates of ~0.75 (although they were lower when computed among the technical students, a first hint that measurement may not be invariant across program context). University students tended to outperform technical students, getting 2.2 and 2.7 additional items correct on average (in QR and CR, respectively). These differences were significant (even small differences would be significant given the sample sizes).

TABLE 1
Descriptive Statistics for 106,189 Spring 2012 SABER PRO Test Takers in Two Assessments

	QR		CR	
Statistic	University	Technical	University	Technical
n	74,997	31,192	74,997	31,192
n, no missing	68,915	28,044	70,290	28,497
M	16.677	14.448	19.996	17.248
SD	6.030	4.839	5.267	4.930
Alpha	0.808	0.695	0.749	0.695
Difference in means	2.229		2.749	
T statistic for difference	60.375		77.832	

Note. Both tests contained 35 multiple-choice items. QR = quantitative reasoning; CR = critical reading.

Figure 1 contextualizes the overall distribution of sum scores for university and technical students (red and blue lines, respectively) and demonstrates that although there are certainly differences in performance between university and technical students, they are not as sizeable as one may suspect a priori. As a point of comparison, the green curves show sum score distributions for university students in the bottom quartile of universities (as measured by mean SABER PRO sum scores; solid line) versus the top quartile (dashed line). In terms of SABER PRO performance, the difference between university and technical students is less than that of students at high- and low-status universities.

Methods

Item Response Theory (IRT)

As with many major assessments, scale scores for the SABER PRO are generated via IRT (e.g., Lord, Novick, & Birnbaum, 1968). If X_{ip} is the Bernoulli random variable

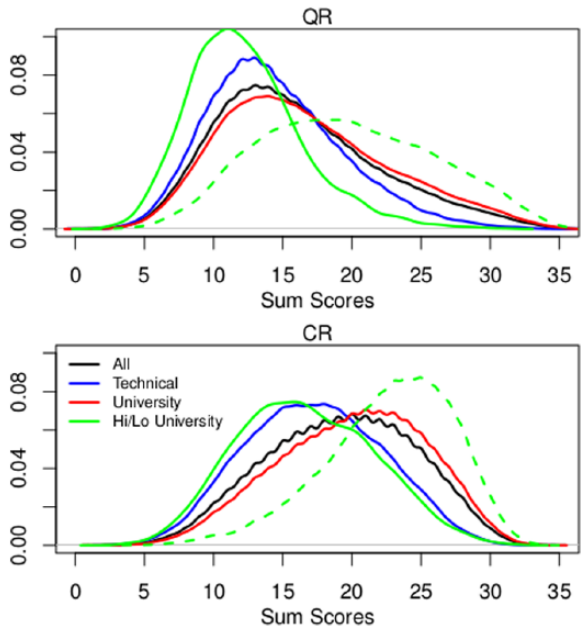

FIGURE 1. *Estimated sum score densities for all students, university and technical students, and students in low- and high-status universities (defined as universities with SABER PRO means in the bottom and top quartiles when computed separately by subject).*

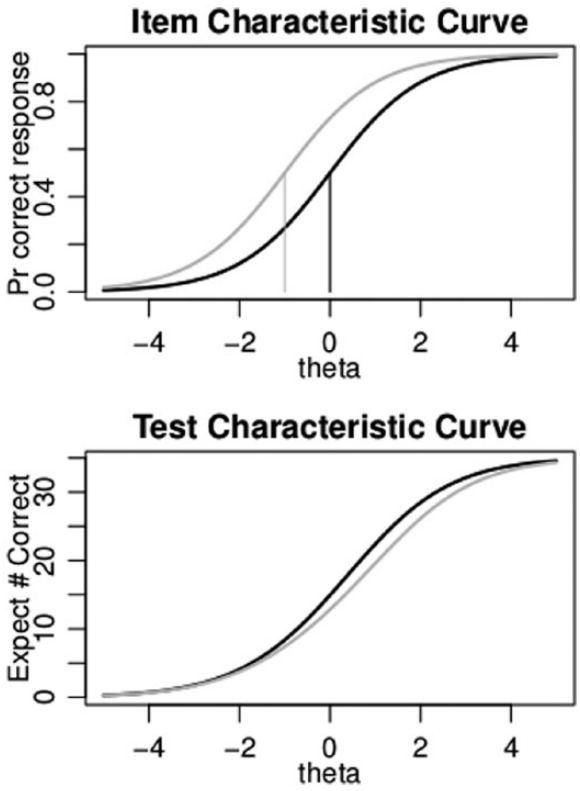

FIGURE 2. *Hypothetical examples of item characteristic and test characteristic curves.*

representing person p's response to item i (correct = 1, incorrect = 0), a standard IRT model (the two-parameter logistic [2PL]) for dichotomously coded items posits that

$$\Pr\left(X_{ip}=1|\theta_p\right)=\frac{\exp\left[a_i\left(\theta_p+b_i\right)\right]}{1+\exp\left[a_i\left(\theta_p+b_i\right)\right]}, \quad (1)$$

where θ_p is an individual's latent ability, b_i is the item easiness, and a_i is the item discrimination. Note that were the individual ability manifest instead of latent, this would be a standard logistic regression problem. We focus on the Rasch model (i.e., $a_i = 1$ for all i) because it is used to scale the SABER PRO in practice. One key assumption we make is about the underlying distribution of θ_p. We assume that abilities for technical school students are distributed as Normal[μ_T, σ_T^2] and university students' abilities are distributed as Normal[μ_U, σ_U^2]. Estimates of μ_T and μ_U can be used to examine the mean difference in ability between university and technical students.

To aid interpretation of our DTF analyses, it helps to further consider the connection between θ_p and the probability of a correct response, $\Pr(X_{ip} = 1|\theta_p)$, given by Equation (1). This relationship can be represented by the item characteristic curve, which describes the probability of correct response to an item as a function of individual ability. Two exemplars are shown in the top of Figure 2. The solid black line shows the item characteristic curve for an item with easiness parameter of 0. The gray line represents an item with easiness parameter of 1. The gray item is easier: Holding ability constant, an examinee is more likely to answer it correctly. Figure 2 can also be interpreted as an example of differential item functioning (DIF; Camilli & Shepard, 1994). DIF occurs when items do not function consistently for different types of students. In our case, suppose the black line represents the item function for a technical student, and the gray line represents the function of the same item for a university student. For the same ability, a technical student is less likely to get the item right (e.g., the item shows evidence of bias).

Item characteristic curves can be aggregated across an entire test to produce test characteristic curves. Two hypothetical test characteristic curves—which show the expected number of correct responses on the entire test as a function of ability, $\sum_i \Pr\left(X_{ip}=1|\theta_p\right)$—are shown in the bottom of Figure 2 (both test characteristic curves describe 35-item tests as is the case with the SABER PRO). Especially at average abilities, examinees are likely to get more of the items from the black test right relative to the gray test. These curves could equivalently characterize different response behavior from two different groups of examinees to the same assessment. The fact that the observed score for a student

with a given ability will vary as a function of group membership is an example of DTF. For example, consider the performance of an English language learner (ELL) on a mathematics test. Compared to other students of similar abilities in mathematics, English learners might be expected to do slightly worse on those items, such as word problems that involve lots of written text. Their performance on those items might suffer due to differences in English language ability compared to their mathematically similar peers. In the following section, we consider methods for detecting DTF and quantifying its impact on estimated ability differences between university and technical students.

Understanding DTF

Equation 1 as written assumes that the item parameters are fixed across student type (university or technical student). There is no reason this need be the case. For example, some items may more closely resemble content from university classes, thus rendering a university student more likely to correctly respond to their items compared to a technical student of similar ability. We consider a model that allows for such group-specific variation in item difficulty:

$$\Pr\left(X_{iTp}=1|\theta_p\right)=\frac{\exp\left[\theta_p+b_{iT}\right]}{1+\exp\left[\theta_p+b_{iT}\right]}, \quad (2)$$

where the item characteristics now depend upon school type *T*. From the item characteristic curve perspective, we are merely allowing for horizontal translations of the item characteristic curve as a function of program type. To estimate Equations (1) and (2), we use the EM algorithm as implemented in Chalmers (2012; a general discussion of the EM algorithm can be found in Do & Batzoglou, 2008, and an IRT specific discussion in Harwell, Baker, & Zwarts, 1988).

Identification of Equation (2) additionally requires the identification of anchor items whose parameters do not vary across institutional context. Doing so is a nontrivial task because results can be sensitive to this choice of anchor items. We utilize a slight variation on a method for anchor item identification outlined previously (Verhagen, Levy, Millsap, & Fox, 2015) and conduct a replication of earlier results. Details are available in the online Supplemental Information (SI). We generally find items to be more comparably functioning across QR (31 of 35 items showed some evidence for invariance compared to only 17 of 35 in CR), but note that even small differences over many items can have effects. Indeed, this is the point of considering DTF. In what follows, we focus on the liberal option of using only a single anchor item, specifically, the item from the SI that showed the most evidence for invariance. We also consider a sensitivity analysis probing the implications of this single anchor item design compared to alternative strategies and return to this topic in the Discussion.

Examining whether items exhibit DIF is typically performed to identify, and then remove, problematic items. In preliminary analyses, we considered the degree to which there was DIF as a function of school type. Using the common "delta" statistic for DIF (Holland & Thayer, 1985), no items showed substantial DIF. However, such DIF analyses do not rule out the possibility of relatively subtle item-level effects that, in the aggregate, may have real impact on our understanding of student ability, the psychometric equivalent of a death by thousand cuts. Our investigation of DTF builds on earlier work (Chalmers et al., 2015) and focuses first on the differences in test characteristic curves for each program type when we estimate Equation (2). In Figure 2, this amounts to a consideration of the vertical distance between black and gray test characteristic curves and tells us where in the ability spectrum DTF is likely to cause the largest problem. We also summarize this information by examining the effect size difference between university and technical students by considering

$$\left(\mu_U-\mu_T\right)\Big/\sigma_U \quad (3)$$

(a variation on Glass's effect size measure; Hedges & Olkin, 2014), based on both Equations (1) and (2).

Results

Measurement Differences Across Program Context

We begin by considering the degree to which item parameters that allow for DTF vary across program context. Figure 3 considers the difference between item easiness parameters across university and technical students (i.e., values higher on the *y*-axis represent items that are easier for university students than technical students conditional on ability) from Equation (2) as a function of the item's easiness when parameters are held fixed across institutional context in Equation (1). The plotted numbers represent the rank of the item in terms of the pseudo *z* statistic (see SI). So, for example, the item with the smallest *z* statistic (1) in each subject was used as the anchor item in applications of Equation (2) and is shown in gray. Note that these items show no difference in easiness between the groups (due to their use as anchor items, this is mechanical). In the top panel, essentially all QR items show some degree of advantage for university students. For CR items (bottom panel), there is less reason for concern regarding DTF as items show bias in both directions.

The left-hand side of Figure 4 provides information on where measurement distortions due to DTF are most pronounced. These figures examine the difference in total sum score (as estimated by Equation [2]) that would be observed for a university student and technical student of common ability (equivalent to the vertical distance between the black and gray lines in Figure 2). Horizontal lines show the overall sum score difference between the groups from Table 1.

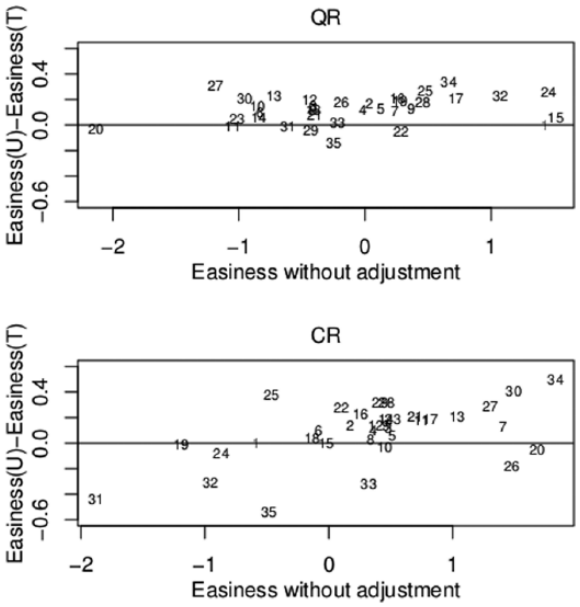

FIGURE 3. *Differences between easiness parameters for university and technical students after adjusting for differential test functioning as a function of the unadjusted easiness estimate. Numbers represent the ordering of items from most to least evidence regarding invariance (see Table S1 in the online Supplemental Information). Gray values are the identified anchor items.*

Scores at the extremes are largely unaffected by DTF, whereas those with abilities in the middle may be affected by about one item (less in the case of CR). These differences represent measurement distortions that are a meaningful percentage of the mean sum score difference for those near the center of the scale (where the majority of the students will typically be located).

Impact of DTF on Perceived Difference in Ability of University and Technical Students

We now consider the extent to which our understanding of differences in the abilities of university and technical students may be biased due to the presence of DTF. Due to evidence shown in Figure 3, we expect to see smaller changes after adjustment in CR because there is less systematic bias in favor of university students in the item parameter estimates. Estimates related to group ability are shown in Table 2. The right-hand side of Figure 4 focuses on the impact of DTF from the perspective of effect sizes (Equation [3]). The black bars show the effect size difference between university and technical students when the test is assumed to function consistently across both groups (e.g., group means and standard deviations come from an application of Equation [1]). The effect sizes are 0.39 and 0.58 for QR and CR, respectively. The gray bars show the estimated effect

TABLE 2
Group Mean Estimates Before Adjustment for DTF (Equation [1]) and After DTF Adjustment (Equation [2])

	QR		CR	
Student Group	μ	σ^2	μ	σ^2
No adjustment				
University	0.00	0.60	0.00	0.41
Technical	−0.30	0.32	−0.38	0.32
DTF adjustment				
University	0.00	0.60	0.00	0.42
Technical	−0.17	0.32	−0.29	0.31

Note. DTF = differential test functioning; QR = quantitative reasoning; CR = critical reading.

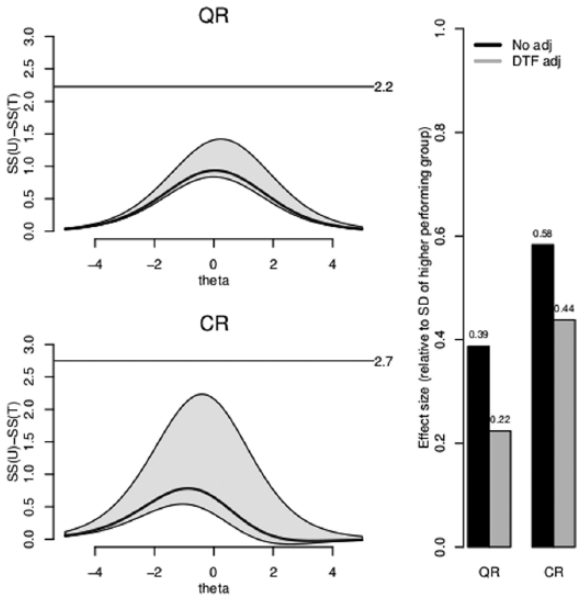

FIGURE 4. *Estimated effect of differential test functioning (DTF) on performance difference on the SABER PRO between universities and technical programs. On left, difference in sum scores (derived from differences in test characteristic curves) as a function of theta after adjustment for DTF. Horizontal lines represent mean sum score differences between school types. On right, estimated effect size difference between ability distributions before and after adjustment for DTF. We generated 95% confidence intervals via 100 bootstrap iterations (where we repeatedly resample from the total pool of students).*

size when DTF is accounted for by allowing item parameters to vary (for all but one item) across group in Equation (2). Effect sizes here are reduced in magnitude to 0.22 and 0.44, respectively. These are substantial reductions, of 44% and 24% relative to the baseline effect sizes, respectively, for QR and CR and may communicate distinctly different stories

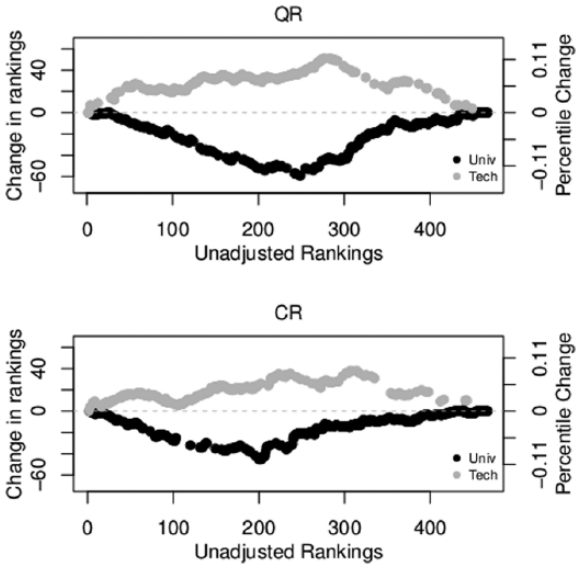

FIGURE 5. *Changes in program rankings before and after adjustment for differential test functioning.*

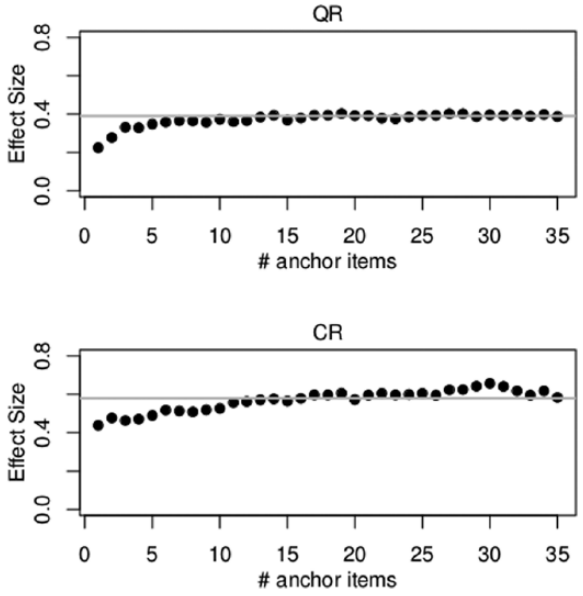

FIGURE 6. *Distribution of effect size difference after adjustment for differential test functioning based on increasing the sample of anchor items. Gray line represents the difference before adjustment. Items are added to set of anchor items as ordered by their associated evidence regarding invariance (i.e., the pseudo z statistic described in the online Supplemental Information).*

about the relative ability of mean university and technical students at the end of higher education.

We also examined the extent to which DTF may affect our understanding of the ordering of program means (N= 467, of which 192 are technical programs). Technical school means of student ability (computed using Warm's [1989] WLE estimator) move up by 0.12 and 0.09 (in QR and CR). University means move down slightly but change much less (<.001 in magnitude across both subjects). In Colombia, for evaluatory purposes, there has been considerable interest in rankings based on the means (e.g., Bogoya, 2012). Figure 5 considers the degree to which rankings may be sensitive to these changes. The technical programs are shown in gray. Individual programs showed substantial movement (see the right-hand side of Figure 5). The mean technical program moved up nearly 29 places on the QR test and 19 places on the CR test. These mean changes in the rankings correspond to percentile changes of 4 to 6 points.

Sensitivity Analyses

To probe the sensitivity of findings to the choice of anchor item, we also conducted analyses where we allowed increasing numbers of items to act as anchor items. Figure 6 shows the effect sizes conditional on these increasingly large sets of anchor items. The sets were constructed so that, relative to the previous analysis, the next item added to the set of anchor items was the item remaining in the pool of items being allowed to vary across program type that showed the most evidence for invariance (as per the pseudo z statistics from Table S1 in SI). The gray line denotes the effect sizes before adjustment for DTF. For QR, the effect size goes from roughly 0.2 to 0.4 after the introduction of five anchor items. That is, the adjustment for DTF is essentially minimal after we are holding constant parameters of the first five items. Based on the top of Figure 3, this is not surprising given that these items (numbers 1–5) all show some small degree of bias toward university students under the single anchor item design. Once they are being forced to function equally across the two groups, we have mechanically removed the impact of DTF. For CR, a return to the unadjusted effect size is more gradual. We return to a discussion of the implications of these findings in the Discussion.

We also use information from four specific areas of study (business, art, engineering, health) that had students at both types of institutions (sample sizes are in Table 3) to further probe the sensitivity of inferences regarding group means to DTF. Figure 7 focuses on two sets of effect sizes. The comparisons in red are between universities and technical schools where the students have the same area of study. As we would expect, university students (used as the reference in Equation [3]) are always performing better than technical school students in the common area of study. The comparisons in black are between university students in different areas of study (for these comparisons, the group with more students is used as the reference group for the purposes of computing effect sizes using Equation [3]). Note that these comparisons fall largely along the 45-degree line, suggesting little evidence for DTF (at least as a function of the

TABLE 3
Sample Sizes for Different Areas of Study

Study Area	University	Technical
Business	22,290	14,160
Art	2,140	2,976
Engineering	14,530	5,456
Health	4,815	2,191

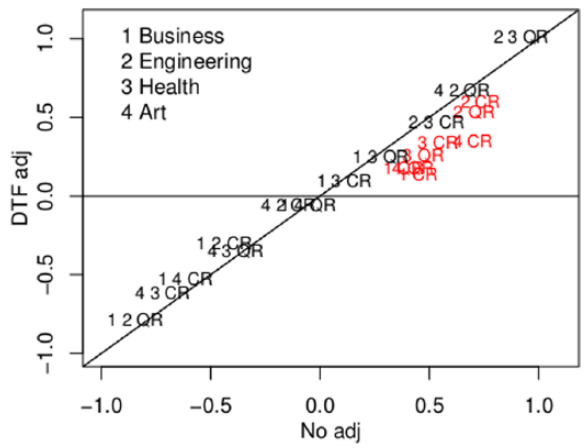

FIGURE 7. *A comparison of effect sizes based on various areas of study before and after adjustment for differential test functioning. Effect sizes in red are between technical students and university students studying similar subjects (e.g., 1 QR shows the estimated effect size difference between business programs in universities and technical schools in quantitative reasoning [QR]), whereas effect sizes in black are between university students studying different subjects (e.g., 4 3 QR shows the effect size difference for health and art students in universities in QR). The 45-degree line is also shown.*

anchor item chosen here) in these cases. The comparison between university and technical students, on the other hand, shows some clear evidence for the relevance of DTF adjustments (i.e., the red estimates are uniformly below the 45-degree line).

Discussion

As attention has increased both domestically and internationally on measuring learning in higher education, the measurement of learning outcomes in technical schools should not be ignored. This admonition is motivated by equity concerns as, compared to university students, students at technical programs are more likely to be historically underrepresented minorities and those of lower SES, but it is also a potentially important part of monitoring the sector for poorly functioning colleges. We examine the possibility of using a measure of learning designed for both university students and technical school students. This needs to be done carefully as higher education exposes students to a potentially much broader array of pedagogical environments than do primary and secondary schools. That said, there are still universal skills that higher education should provide students (e.g., increases in their ability to think critically and engage with technical material), and the key question here is whether a measurement instrument designed for university students can be used equally well with technical school students.

If the goal is to separately analyze university and technical school students, the SABER PRO would likely perform adequately in either group although the SABER PRO has slightly reduced reliability among technical students compared to the reliability among university students. However, things become more complicated when the SABER PRO is scaled jointly with university and technical students. Despite the fact that the items do not exhibit substantial DIF, there is evidence for DIF's perhaps more sinister generalization, DTF. The existence of DTF has implications for our measure of mean differences (Figure 4), reducing the observed effect size difference between the two groups by up to 44% in QR and 24% in CR. These differences translate into potential differences in how schools are ranked on outcomes (Figure 5). DTF can mean that a school's rank ordering changes by an entire decile in some cases, a highly important fact given that the SABER PRO is widely used for the purposes of school ranking and the evaluation of institutions of higher learning. Moreover, there is evidence (Figure 7) to suggest that these findings are due to something unique about the difference between technical and university programs.

On the Uses of DTF

From our perspective, DTF is an underutilized concept whose time has come. Considerations of measurement variation (e.g., the literature on DIF) largely focus on individual items. Although such questions are reasonable starting points, we think that asking about the overall implications of even a small amount of measurement variation should be a standard part of psychometric practice. As we have observed here, performance differences between groups of students can be substantially biased by many small item biases. This is particularly relevant in the age of value-added analyses wherein teachers and schools are judged on the basis of test scores, which may in some cases be influenced by DTF along the lines we describe. Others have examined how issues of scale construction affect value-added type analyses (Briggs & Weeks, 2009), but none have directly addressed the issue of DTF to our knowledge. Computational limitations have perhaps been one reason that DTF has not played a major role in educational measurement up to this point, but with the development of new computational tools (Chalmers, 2012; Verhagen et al., 2015), there is room for this to change.

We note an interesting difference between QR and CR with respect to DTF. In the SI, more items show evidence for

differential functioning for CR, and yet we observe in Figure 4 a smaller effect of DTF on CR. Why is this? We suspect it is due to the nature of the item-level differences in CR that are shown in Figure 3, which suggests that CR items, although less likely to function consistently across groups, show some degree of balance in this inconsistency (some favor university students; some favor technical students). QR items, on the other hand, uniformly favor university students. Even if the individual items perform better across group, as a whole they favor university students, leading to the larger impact of DTF on the QR exam. It should be noted that evidence from Figure 3 is based on our use of a single anchor item. This is a topic to which we now turn.

Figure 6 probes the sensitivity of our main findings regarding effect sizes to the inclusion of additional anchor items. We focus this discussion on the QR exam. As previously discussed, the effect of DTF on this test fades after the inclusion of about five anchor items. One potential critique of our findings is that they are contingent upon the identification and use of only a single anchor item. Regarding identification of that item, we have conducted a thorough investigation on this point (see SI) and feel confident that our choice of anchor item is reasonable. Regarding the use of a single anchor item, future research should examine this question in greater detail. Our perspective is that because DTF is typically ignored, research has been based upon the overly restrictive assumption that all items function equivalently across groups. Given the history of this strongly restrictive assumption, we argue that a reasonable starting point for research in this area is to examine DTF using the least restrictive assumptions. That is what we do here, but we encourage others to consider this issue in more detail as research in this domain accumulates.

Although we consider a single assessment, our findings have potential implications for other assessments used in both higher education and other settings. ELL students, for example, are perhaps less likely to answer many test items correctly than their similarly able peers merely as a function of their ELL status. It is these cases, where small biases may be expected to largely fall in one direction, where DTF is likely to be especially concerning. In other cases, specifically those wherein DIF studies show similar numbers and magnitudes of item biases in both directions (as with the CR assessment), DTF is likely to be minimal. Consider the SAT. Evidence for DIF on the SAT as a function of race (Santelices & Wilson, 2010, Figure 1) suggests that there is unlikely to be substantial race-specific DTF for that version of the test (although a more thorough analysis on this point would clearly be of interest).

DTF also offers a potentially promising approach for examining thorny issues of fairness. Returning to the issue of race and the SAT, an earlier debate (Dorans, 2004; Freedle, 2003; Santelices & Wilson, 2010) involved arguments about DIF as a function of item difficulty and associated implications for understanding differences in student performance. At issue was whether associations between DIF and item difficulty were indeed problematic from the perspective of fairness as easy items were found to be harder for minority students. As an isolated fact, this is potentially concerning but its impacts are uncertain. A reframing of this issue in terms of DTF would return the debate to more concrete issues—differences between test characteristic curves at different places on the ability scale—that have potential resolutions using standard tools from the psychometric tool kit (i.e., they would not require the rescaling regressions proposed by Freedle, 2003).

Limitations and Future Research

There are several findings from this study that may not generalize beyond the Colombian context. First, the fact that technical school students are frequently enrolled in the same institutions as university students (but simply taking different tracks) is not typical of the U.S. higher education system, in which those two types of programs are serviced by different institutions. Second, we may be underestimating the effect size difference between university and technical students due to our use of data from the spring test administration, which potentially contain fewer private university students than does Colombia as a whole given that elite Colombian high schools run September to May, whereas most public high schools run January to December. It is also the case that Colombian higher education students take all classes within their area of study as opposed to the more diverse sets of courses typically taken by students in the United States. The major limitation of the present study, and a clear target for future research, would be a comparison of performance at the end of higher education compared to the beginning for both university and technical students. Such a study, looking at differences in pseudogrowth (see Briggs & Domingue, 2013, for a discussion of growth measurement with and without vertical scales) would offer an interesting comparison of the amount that students learn in universities relative to technical schools.

One useful avenue of inquiry for future research would be an investigation into whether there are identifiable item features that are leading to DTF. Such investigations will be expansions on study of item bias in which generalized features of items are potentially leading to DTF (even if individual items do not exhibit DIF). Elements of item difficulty modeling (Gorin & Embretson, 2006; Stenner, Smith, & Burdick, 1983) may be useful in this quest. For example, with the CR test, it may be interesting to evaluate each item's content on a continuum of abstract to concreteness and examine whether this is associated with the observed biases. Finding that abstract items tend to favor university students may offer an avenue for improving the validity of the SABER PRO because such an issue could potentially be addressed. DIF studies are also known to be sensitive to true

differences in ability (e.g., DeMars, 2010), and although evidence from Figure 7 suggests that this may not be an issue with DTF, more evidence on this point would be useful.

Conclusions

In current conceptions, the validity of a test is contingent upon its proposed uses (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, Standard 1.1). Given the interest in Colombia in using the SABER PRO as a metric for institutional effectiveness, examining the sensitivity of ranking based on the SABER PRO to alternative assumptions should be a key component of the validity argument used to support the test. We find that rankings of schools perhaps undervalue the performance of technical schools. Because students may make decisions about where to pursue higher education based on such rankings, this is no small matter.

How to resolve this problem is less clear. At a minimum, students from technical schools (and other lower-status institutions) should be included in the piloting of items designed for higher educational measures if they are a part of the population of interest. Where possible, differences in the content (e.g., do students encounter quantitative reasoning in the form of abstract problems from physics or concrete problems involving daily finance?) encountered by students in these two types of institutions should be considered in the preparation of items. Although addressing these issues will cost time and money, we think they are a crucial part of building next-generation assessments of learning for use in higher education. Given the quantity of students attending technical schools and the importance of higher education for securing a spot in the modern workforce, we think aiming to include these schools as a part of such next-generation assessment systems is a crucial task.

Acknowledgment

The authors would like to thank ICFES for their cooperation during the course of this research.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bogoya, D. (2012). *Elementos de calidad de la educación superior en Colombia* [Quality elements of higher education in Colombia]. Retrieved from https://sites.google.com/a/unal.edu.co/danielbogoya/5-benchmarking-de-universidades

Briggs, D. C., & Domingue, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics*, *38*(6), 551–576.

Briggs, D. C., & Weeks, J. P. (2009). The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy*, *4*(4), 384–414.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

Chalmers, R. P., Counsell, A., & Flora, D. B. (2015). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, *76*(1), 114–140.

DeMars, C. E. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement*, *70*(6), 961–972. doi.org/10.1177/0013164410366691

Deming, D. J., Goldin, C., & Katz, L. F. (2011). *The for-profit post-secondary school sector: Nimble critters or agile predators?* Cambridge, MA: National Bureau of Economic Research.

Do, C. B., & Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature Biotechnology*, *26*(8), 897–899.

Dorans, N. (2004). Further comment: Freedle's Table 2. Fact or fiction? *Harvard Educational Review*, *74*(1), 62–73.

Freedle, R. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review*, *73*(1), 1–43.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*(5), 394–411.

Hambur, S., Rowe, K., & Luc, L. (2002). *Graduate Skills Assessment*. Canberra, Australia: Commonwealth of Australia.

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational and Behavioral Statistics*, *13*(3), 243–271.

Hedges, L. V., & Olkin, I. (2014). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Holland, P., & Thayer, D. (1985). *An alternative definition of the ETS delta scale of item difficulty*. Princeton, NJ: Educational Testing Service.

Instituto Colombiano para la Evaluación de la Educación. (n.d.). *Reporte aporte relative* [Value-added report]. Retrieved from http://www.icfes.gov.co/instituciones-educativas2/saber-pro/resultados-ies/reporte-aporte-relativo

Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The collegiate learning assessment: Facts and fantasies. *Evaluation Review*, *31*(5), 415–439.

Las instituciones con las mejores pruebas Saber Pro [Institutions with the best results in the Saber Pro tests]. (2013, September 12). *El Tiempo*. Retrieved from http://www.eltiempo.com/archivo/documento/CMS-13059672

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Boston, MA: Addison-Wesley.

Los Andes, primera en Pruebas Saber Pro [The Andes, first in Saber Pro tests]. (2013, September 12). *El Nuevo Siglo*. Retrieved from http://www.elnuevosiglo.com.co/articulos/9-2013-los-andes-primera-en-pruebas-saber-pro

Ministerio de Educación Nacional (n.d.). Modelo De Indicadores Del Desempeno De La Educación: Metodologia [Model of

performance indicadors of education: Methodology]. Retrieved from http://www.colombiaaprende.edu.co/html/micrositios/1752/articles-351671_Metodologia.pdf

Milla, J., San Martin, E., & Van Bellegem, S. (2016). Higher education value added using multiple outcomes. *Journal of Educational Measurement*, *53*(3), 368–400.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20.

Santelices, M. V., & Wilson, M. (2010). Unfair treatment? The case of Freedle, the SAT, and the standardization approach to differential item functioning. *Harvard Educational Review*, *80*(1), 106–134.

Shavelson, R., Domingue, B., Marino, J., Mantilla, A., Morales, J., & Wiley, E. (2016). On the practices and challenges of measuring higher education value added: The case of Colombia. *Assessment & Evaluation in Higher Education*, *41*(5), 695–720.

Stenner, A. J., Smith, M., & Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, *20*(4), 305–316.

Tremblay, K., Lalancette, D., & Roseveare, D. (2012). *Assessment of higher education learning outcomes: Feasibility study report, volume 1 design and implementation*. Paris, France: Organisation for Economic Co-operation and Development.

Uniandes, la mejor en Pruebas Saber Pro 2012 [Uniandes, the best in Saber Pro tests 2012]. (2013, September 11). *El Espectador*. Retrieved from http://www.elespectador.com/noticias/educacion/uniandes-mejor-pruebas-saber-pro-2012-articulo-445754

Verhagen, J., Levy, R., Millsap, R. E., & Fox, J.-P. (2015). Evaluating evidence for invariant items: A Bayes factor applied to testing measurement invariance in IRT models. *Journal of Mathematical Psychology*. Advance online publication.

Verhine, R. E., Dantas, L. M. V., & Soares, J. F. (2006). From the National Course Exam (Provão) to ENADE: A comparative analysis of national exams used in Brazilian high school. *Ensaio: Avaliação E Políticas Públicas Em Educação*, *14*(52), 291–310.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. doi.org/10.1007/BF02294627

Zlatkin-Troitschanskaia, O., Pant, H. A., Kuhn, C., Toepper, M., & Lautenbach, C. (in press). Assessment practices in higher education and results of the German research program modeling and measuring competencies in higher education (KoKoHs). *Journal Research & Practice in Assessment*.

Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Studies in Higher Education*, *40*(3), 393–411.

Authors

BENJAMIN W. DOMINGUE is an assistant professor at the Stanford Graduate School of Education. He is interested in psychometrics and sociogenomics.

DAVID LANG is a doctoral student in the Economics of Education program and an Institute of Education Sciences Fellow. His research interests include higher education, online education, and quantitative methods in education research.

MARTHA CUEVAS worked at the Colombian Institute for Assessment Education (ICFES). Her research interests include psychometric and quantitative methods in psychology and education.

MELISA CASTELLANOS is a PhD candidate in experimental psychology, Concordia University (Montreal, Quebec). Her research interests include developmental psychology, education measurement, and peer relationships in childhood and adolescence.

CAROLINA LOPERA works at the Education Evaluation Center at Los Andes University. She is interested in instrument development and validation, program evaluation, and quantitative analysis of educational data.

JULIÁN P. MARIÑO heads the Educational Evaluation Center at Los Andes Education School. His interests include psychometrics, education measurement, and test development.

ADRIANA MOLINA works at the Education Evaluation Center at Los Andes University. She is currently building a system for teacher support and evaluation in higher education.

RICHARD J. SHAVELSON is the Margaret Jacks Professor of Education (Emeritus) and the I. J. Quillen Dean of the Graduate School of Education (Emeritus) at Stanford University. His research focuses on educational and psychological measurement and statistical modeling as applied in practice and policy.