THIS DOCUMENT CONTAINS THE AUTHOR'S ACCEPTED MANUSCRIPT. THE PUBLISHED MANUSCRIPT (ACCEPTED FOR PUBLICATION IN THE JOURNAL OF EXPERIMENTAL EDUCATION ON MARCH 31, 2023) IS AVAILABLE FROM TAYLOR & FRANCIS PUBLISHERS:
https://www.tandfonline.com/toc/vjxe20/current

Classroom Assessment and Instructional Modes:

An Exploration of School-level Contextualized Psychometric Challenges

A. Corinne Huggins-Manley

Jing Huang

Jerri-ann Danso

Wei Li

Walter L. Leite

University of Florida

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C160004 to the University of Florida. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Corresponding author: Corinne Huggins-Manley, 1602 Norman Hall, Gainesville, FL, 32611,

352-273-4342, amanley@coe.ufl.edu

Abstract

The global COVID-19 health pandemic caused major interruptions to educational assessment systems, partially due to shifts to remote learning environments, entering the post-COVID educational world into one that is more open to heterogeneity in instructional and assessment modes for secondary students. In addition, in 2020, educational inequities were brought to the forefront of social conscience. The purpose of this study is to empirically explore how contextual (i.e., school-level) race and economic factors may relate to and explain measurement challenges that can arise during shifts to remote learning. We fit a series of multilevel models to explore school-level factors in assessment data alongside psychometric problems of differential item functioning and person fit in classroom assessment measurement models. Our results demonstrate ways in which our project's classroom assessments were impacted by shifts to remote learning, emphasizing the importance of researchers and practitioners evaluating such concerns when seeking validity evidence for interpretation of classroom assessment data.

Keywords: *classroom assessment, differential item functioning, person fit, instructional mode*

Introduction

The global COVID-19 health pandemic upended all aspects of societies around the world, including major interruptions to educational systems. Measurement and assessment within these educational systems were not spared (Harris, 2020). Meanwhile, the death of George Floyd sparked international unrest surrounding critical issues of racial injustice in the United States and beyond. Educational systems were forced to question the realities of their roles in perpetuating inequality in our society, with a spotlight on educational measurement for its relative lack of attention to the varied experiences of students in the systems (Dixon-Román, 2020; Randall, 2021; Sireci, 2021). Ultimately, the COVID pandemic and the educational inequalities it highlighted call into question much of what we know and do in educational assessment. While schools are now re-opened, the effects of the health pandemic are still felt in many ways, including more openness to remote learning and assessment as well as continued calls for attention to equity and fairness in assessment.

The initial pandemic-related influence on education and assessment systems in the United States occurred in March of 2020. Public K-12 schools around the nation were closed for in-person instruction, sending millions of children into their homes for their school days (García & Weiss, 2020). Meanwhile, the educational impacts of such abrupt and massive changes were inequitably dispersed across racial and economic lines (García & Weiss, 2020; U. S. Department of Education, Office of Civil Rights, 2021). And while the world assumed for a moment that we would shift back to pre-COVID education models within some timeframe, it became clear that some of the COVID conditions have changed many of our classroom conditions for the foreseeable future (U. S. Department of Education, Office of Civil Rights, 2021).

With the massive heterogeneity in instructional delivery during the initial school shutdown, many questions arose as to how to interpret the data coming from classroom-based assessments. With the lack of consistent in-person instruction, would students engage differently with classroom assessments? And vice-versa, would classroom assessments operate differently, psychometrically speaking, without the same in-person classroom instruction conditions under which they were developed? Ultimately, the COVID pandemic and call for action against educational inequality induced many unanswered questions about how classroom assessment data can be interpreted and used validly in these new schooling environments.

The purpose of this study is to empirically explore how contextual (i.e., school-level) factors of race and income may relate to and explain measurement challenges that may have arisen during the substantial schooling changes introduced by COVID. Specifically, we explore psychometric features of algebra classroom assessments in Florida that allow for valid inferences about student knowledge, and how those features may have changed when schools and testing systems shifted into the COVID and post-COVID era. We chose to investigate the phenomena in algebra because this study was conducted within a larger project focused on the teaching and learning of algebra in a virtual learning environment. We relate these psychometric changes to school-level race and economic factors that are known to associate with academic outcomes (as measured by assessment scores) in educational systems, including inequities experienced during the school closures (García & Weiss, 2020; U. S. Department of Education, Office of Civil Rights, 2021).

To achieve this purpose, we explored differential item functioning and person fit under a two-parameter logistic measurement model framework (2PL; Birnbaum, 1968), modeled

concurrently with school-level explanatory variables related to racial and economic contextual conditions. The research questions are:

1. To what extent did algebra classroom assessment items display differential item functioning across pre-pandemic and pandemic groups of student data?
2. To what extent did algebra classroom assessment item models display differences in person fit across pre-pandemic and pandemic groups of student data?
3. How are the differential item functioning and person fit findings across pre-pandemic and pandemic groups of student data related to school-level race and economic variables?

This study is couched within a larger research project that was collecting data during the spring of 2020 through a virtual environment, Math Nation (Lastinger Center for Learning, 2019). The assessment score inferences on the larger project were supported by multiple pieces of validity evidence under the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), but that evidence was gathered before the COVID pandemic (Xue et al., 2022). To continuously evaluate algebra learning in this project, despite such massive disruptions to schooling and testing, it was critical to establish that the underlying measurement models were continuing to represent the student data and that interpretations of such assessment data were made in consideration of relationships to school-level race and economic factors.

In sharing our explorations into this phenomenon, we align with the notion that researchers must continue to provide empirical evidence of how shifts to remote or hybrid instruction and assessment has impacted education and, importantly, inequalities in educational

assessment experiences (U. S. Department of Education, Office of Civil Rights, 2021). Specifically, we intend to share our findings to contribute to the literature base that assists educators in understanding ways in which inequalities might predict classroom measurement phenomena. In addition, we intend to model to the field some critical psychometric work that can be done with classroom assessment data to monitor, understand, and interpret assessment scores in education systems dealing with heterogeneity in mode of instruction (i.e., in-person vs. remote) and the presence of racial and economic differences across schools.

Psychometric Framework

There are many psychometric features expected of measurement data to support valid interpretations of student assessment scores (AERA, APA, & NCME, 2014). These expectations extend to classroom assessment initiatives, as a global aim of such initiatives is to infer valid information about student abilities. When classroom assessments are undergirded by a psychometric model, such as an item response theory (IRT) model, two of many possible psychometric considerations are that of measurement invariance and model fit to data. Within those two considerations are some specific statistical methods for evaluating the psychometric properties of interest for internal structure validity evidence. In this study, we focus on methods for detecting differential item functioning (DIF) and person fit.

DIF has been a widely studied psychometric phenomenon in large scale testing and educational measurement research for over half a century (Brennan, 2006). DIF is defined as the presence of group differences in item responses after conditioning on the trait of measurement (Penfield & Camilli, 2006). In achievement tests intended to measure one trait with binary scored items, DIF can be more specifically defined as the presence of group differences in the probability of obtaining a correct response after conditioning on the single latent trait of

measurement. While the presence of DIF may or may not indicate item bias, DIF does indicate that the parameters of the underlying measurement model differ between groups (Penfield & Camilli, 2006). In this study, we use DIF analysis to understand if item difficulty was invariant (i.e., consistent) across the pre-pandemic and pandemic schooling environments. A lack of invariant measurement would indicate that the assessment was measuring at least one secondary unintended trait within or across the two time points of study (Penfield & Camilli, 2006). This type of result would call into question the validity of any inferences drawn from such assessment scores. Given that we are defining our groups by time in this study, one might view the DIF analysis in this study as a form of item parameter drift analysis (Goldstein, 1983).

Person fit lies in the larger field of model-data fit in parametric statistics and has been a studied feature in IRT model applications since at least the 1990s (Brennan, 2006). Presence of person fit refers to consistency between an examinee's response pattern on the assessment and the expectations stemming from the psychometric model (Embretson & Reise, 2000). Person fit indices have generally been used to identify unique cognitive processes of test examinees as well as aberrant testing behaviors (Brennan, 2006; Karabatsos, 2010). Regardless of why a model might fail to capture the underlying patterns of a student's item response data, a lack of person fit for a particular student statistically indicates that the measurement model is not a good representation of the student's assessment data, which calls into question the validity of any inferences drawn about that student from that measurement model (Brennan, 2006). Hence, in classroom assessment, and more particularly this study, problematic person fit results indicate a need to understand more deeply the student response patterns that are not captured by the model and hence cannot be inferred upon in the same way as other students. Given that we are defining our groups in this study as pre-pandemic assessment takers and pandemic assessment takers,

differences in person fit across these groups indicates that the validity support for inferences about student algebra knowledge after students were sent to learn and assess in remote environments is not consistent with the support that we had for such inferences prior to school closures.

While it is established that the presence of DIF or the presence of person misfit can threaten the validity of inferences drawn from classroom assessment scores, it is often difficult to determine why such measurement problems are present. Without such knowledge, it is hard to go from the statistical interpretation of DIF or person misfit results to a meaningful understanding of why and for whom assessments are not producing the desired results about all test takers. While explaining the role that school-level racial and economic factors play in DIF and person misfit does not provide a true "why" to the psychometric problems, it does help us to understand if the impacts of such problems are experienced differently for different school contexts related to student race and income, which can help us to understand important connections between measurement problems and inequality. In this study, we hypothesize that the introduction of the health pandemic and its associated shift to remote assessment might have resulted in psychometric problems in classroom assessment, preventing useful and valid interpretations of student knowledge from such assessments, and that these issues may be partially explained by the racial and income compositions of the student populations served by schools.

Recent methodological research has encouraged and proposed methods for the examination of contextual factors in schooling that may help to explain psychometric findings, such as DIF (e.g., Vo & French, 2021). In the context of DIF, such contextual examinations are pertinent due to what we have learned from a long history of evaluating DIF, namely that it is no longer as difficult to detect DIF as it is to understand why it has occurred or what it means for the

larger context of education and learning (Zumbo, 2007). A similar argument could be made for person fit explorations, in which there is a need to go beyond identifying misfitting persons with statistical indices to understanding under what contexts person misfitting may occur. Using methods to evaluate contextual factors related to DIF and person fit, we are able to explore how contextual factors related to the pandemic may explain psychometric issues that may invalidate, or at least help us to further contextualize and understand, our classroom assessment interpretations and uses.

Data Sources

Math Nation (Lastinger Center for Learning, 2019) is an educational technology platform containing multiple sets of curriculum units offered to teachers in multiple states, including Florida. The algebra portion of Math Nation has experienced widespread adoption in Florida schools, with over 6,000 teachers and 250,000 students accessing materials on an annual basis. The video portions of the curriculum are aligned to the Florida State Standards, with practice assessments aligned tightly to the End-of-Course high stakes algebra exam in the State. Passing this test is a graduation requirement for Florida public high schoolers, however that requirement was waived in the spring of 2020 when the test was not administered (U.S. Department of Education, 2020a).

In the spring of 2020, our larger project introduced a randomized control trial related to algebra learning in three public school districts in Florida. The teachers were incentivized to have students complete multiple videos and accompanying practice assessments each week. When these schools were closed due to the COVID-19 pandemic, usage on the Math Nation platform increased, as expected because the virtual nature of the curriculum was appealing once teachers were required to teach students remotely. Despite the state End-of-Course algebra exam

being cancelled soon after school closures (U.S. Department of Education, 2020a, 2020b), we did not observe a decrease in usage of Math Nation, but rather an increase. However, we recognized the need to evaluate the psychometric properties of the practice assessments, given changes in administration conditions (i.e., taking most practice assessments outside of school) and the potential for unequal learning and assessment conditions in home and school environments divided by racial and economic lines.

To explore these psychometric properties, we selected two sections of the Math Nation algebra curriculum: Exponential Functions and One-Variable Statistics. As Florida teachers are generally following the same sequence of algebra curriculum in their classrooms (i.e., the sequence aligned with the Florida State Standards), we selected sections that were aligned with the timing of the school closures and, hence, where many teachers and students in the state were providing relatively large amounts of data across the pre-pandemic and pandemic groups on the same assessments, which was required for multilevel and IRT modeling. The Exponential Functions assessment consisted of 18 items, and the One-Variable Statistics assessment consisted of 27 items. Like the statewide End-of-Course Algebra exam in Florida, the items were a mix of multiple choice, fill-in-the-equation (with an equation editor), mutli-select items, and more (see Florida Statewide Assessments Algebra 1 example items: https://fsassessments.org/resources/). Due to computational burdens in estimating item parameters and real-time ability in the larger project, all items were scored as binary and parameterized under the 2PL model, even for multi-part items (see Xue et al., 2022). Throughout the larger study where students are scored on assessments in Math Nation, we treat the assessments separately. Hence, in this study, we evaluated each research question through a replication of analyses across two assessments.

We used data only from students who completed one or both of the two assessments in the virtual learning platform during the spring 2020 semester. We focused on complete assessments for four reasons. First, as is the case in many virtual learning platforms, the unstructured nature of the data resulted in large amounts of missingness and large amounts of data points with questionable meaning (Cope & Kalantzis, 2016; Huggins-Manley et. al., 2019). Complete assessments in our larger project have often been used to indicate a more teacher-driven and graded assignment of the assessment, and hence a more supported assumption that the logfile data comes from a meaningful and effortful attempt on the student's part to engage in the assessment. This type of teacher-driven assessment, we feel, is more connected to formal classroom assessment, as opposed to students self-selecting to take items on their own for no classroom grade and putting in unknown amounts of effort. This may be particularly relevant in a school year in which some student motivations and efforts in learning Algebra may have changed once the end-of-course graduation testing requirement was waived (U.S. Department of Education, 2020a). Second, the multilevel modeling method we use to evaluate contextual features of psychometric findings in this study introduces some interpretation challenges even under the best of data conditions (Vo & French, 2021). Having complete assessment data that was teacher-driven helps to mitigate, or at least not add to, the already challenging interpretation decisions. Third, missing data imputation is challenging in logfile data, and not enough is known to confirm our decisions and assumptions in such a process with our data and planned analysis. For example, many students who did not complete the assessments completed only a single item, leaving large amounts of missingness, much uncertainty about the motivation of the student while taking the assessment, and no additional data to use as imputation predictors. Finally, given that MathNation is available to students and teachers in an ongoing manner, our data never

really contains “all students” when we do a capture of the data at a given timepoint, which is expected when using data from virtual environments with ongoing log files. For these reasons, we only included complete assessment data from Spring 2020 in our study, which poses some limitations for generalizing our findings to all students studying algebra with Math Nation. We interpret our findings in light of this limitation.

Methods

The Exponential Functions assessment data consisted of 927 students across 27 schools, in which 670 students completed the assessment pre-pandemic (the reference group in the DIF analyses) and 257 students completed the assessment during the pandemic (the focal group in the DIF analyses). The One-Variable Statistics assessment data consisted of 1,019 students across 23 schools, in which 283 students completed the assessment pre-pandemic (the reference group in the DIF analyses) and 736 students completed the assessment during the pandemic (the focal group in the DIF analyses). The unbalanced groups, as well as the reversal of focal-to-reference dominant group sizes, is because majority of teachers had completed the Exponential Functions portion of the curriculum sequencing prior to pandemic-related school closures, whereas majority of teachers had not yet reached or completed the One-Variable Statistics portion of that sequencing. While unbalanced designs can impact type I errors and power in DIF analyses, we believe the effect should be minimal due to all groups having over 200 students and our overall sample size nearing 1,000 students (e.g., Paek & Guo, 2011). Hence, we aimed to include as many students as possible in our analyses despite the unbalanced nature of this observational data. We also provide interpretations around corrected alphas in our methods and results to address Type I error rates, and we provide effect sizes for interpretation of practical effects to consider in the case of non-significant results that may come about due to concerns of power.

We have not included teacher-level information in this sample size description, nor do we include teacher-level variance in our analysis plan, because many schools in our project had one or only a few algebra teachers serving the student body. Hence, school and teacher level concerns are largely confounded in our datasets.

Table 1 shows the item means as well as the point-biserial correlations of each item with the total summated score, all computed within (not across) each assessment. Overall, the items were easier and more discriminating (i.e., positively correlated with the total test score) in the Exponential Functions assessment as compared to the One-Variable Statistics assessment. The differences in difficulty might stem from both the nature of the curriculum sequencing under the Florida State Standards (i.e., One-Variable Statistics is considered a higher-level algebra domain) and the fact that more students in the One-Variable Statistics dataset were learning in the pandemic schooling environment (i.e., schools closed; cancellation of high-stakes algebra test).

School-Level Predictors

All predictors were taken from Florida Department of Education (FDOE) public data, including school district ID (total of three districts), four school-level demographics (for control purposes) and economics variables (i.e., Title 1 status of the school, the percentage of economically disadvantaged students (“Ecodis”), the percentage of English language learners "ELL”, the percentage of students with disability “Disability”), and three race/ethnicity variables (i.e., the percentage of Hispanic students, the percentage of Black students, and the percentage of White students). All other races (e.g., Native Hawaiian or Other Pacific Islander) were excluded due to the small percentage of these students per school. Table 2 shows our sample descriptive statistics of these school-level variables and student test scores, disaggregated by the two assessments and by district.

While we aimed to include all the publicly available school-level predictors, we found many of them to be highly correlated in our database. Hence we conducted the following variable selection procedures to pick the best set of predictors for the multilevel DIF analysis: (1) checked the correlations among school-level variables; (2) ran school-level OLS regressions using all school-level variables as independent variables and using the aggregated probability of the correct response to each DIF items (i.e., at the school-level) as the dependent variable, and then checked the multicollinearity with Variance Inflation Factor (VIF); (3) ran factor analysis with all school-level variables.

Table 3 and Table 4 show that some school-level race variables were highly correlated with some multicollinearity issues when entered with other predictors, and hence we could not include all three race variables in one multilevel model. In addition, the tables show that "Title 1" and "Ecodis" were similar and correlated. According to the findings of the factor analyses, "ELL" and "Disability" can measure something different than "Title 1" and race variables, but they both lacked enough variability across schools and were not statistically significant when we included them in the final multilevel models. Considering this, the set of variables we used as the final school-level predictors include school district dummy variables, Title 1 status of the school, and the race variable "percentage of Black students."

Differential Item Functioning

To explore DIF and the contextual factors surrounding it, we fit a series of multilevel logistic regression models (Chen & Zumbo, 2017; Vo & French, 2021) including predictors at the student level (level 1) that assist in evaluating uniform DIF as well as the selected contextual predictors at the school level (level 2).

We first identified a set of items that were DIF-free in a single level context. These items could then serve as a set of DIF-free items to calculate a purified total test score, which could then be incorporated into the multilevel models used to evaluate contextual factors of DIF for each of the remaining items. We used logistic regression DIF methods in the difR package (Magis, Beland, & Raiche, 2020) in the R environment (R Development Core Team, 2020), with a purification method to select a final set of DIF-free items. The purification method follows findings of Clauser and Mazor (1998) in which items are iteratively tested for DIF, allowing different configurations of assumptions about which items are DIF-free, until multiple iterations indicate the same set of DIF-free items (Magis et al., 2020). After purification, a final set of target items was identified for contextual evaluation of DIF, while the DIF-free items were used to develop a purified total test score to condition the group differences in the multilevel models on the measured trait, thereby allowing uniform DIF detection.

We used Stata (StataCorp, 2021) with maximum likelihood estimation for all DIF-related multilevel analyses in this study. With each target item as the dependent variable in a series of models, we began the analysis process by fitting a single level model including only the purified total test score and the binary grouping variable (pre-pandemic = 0, pandemic = 1) as predictors. We used a statistically significant odds ratio at the $\alpha = .05$ level for the grouping variable to indicate the presence of uniform DIF. We expected these results to align with those in the difR package, but were open to small differences, given differences in estimation between Stata and difR as well as the fact that the Stata logistic regression analysis assumed a fixed set of purified items for the conditioning variable.

We then added schools as level 2 clusters with no school-level predictors to explore if some of the DIF effect could be explained by school-level variance that was ignored in the single-level model.

Next, we added school Title 1 status and the percentage of Black students as predictor variables at the school-level, along with dummy variables for the three different districts to which the schools belonged, to explore how these predictors might relate to some of the DIF effects while controlling for district differences. After adding all predictors to the school level, and retaining the purified total test score (PurifiedScore) and the binary grouping variable (Pandemic) at level 1, the final random-intercepts multilevel logistic regression models used to evaluate DIF in each target item were specified as

$$\eta_{ij} = \gamma_{00} + \gamma_{10}(PurifiedScore_{ij}) + \gamma_{20}(Pandemic_{ij}) + \gamma_{01}(Race_j) + \gamma_{02}(Title1_j) + \gamma_{03}(District\_Dummy_{j1}) + \gamma_{04}(District\_Dummy_{j2}) + u_{0j}, \quad (1)$$

where $\eta_{ij} = \ln(\frac{P(Y_{ij})}{1-P(Y_{ij})})$ in which Y is an item response; i represents an item; j represents a school; γ_{00} is the conditional average log-odds of item response across schools; and $District\_Dummy_{j1}$ and $District\_Dummy_{j2}$ represent two district dummy variables. We then added an interaction term between the Title 1 variable and the Percent Black variable, hoping to explore this one type of intersectionality. However, the interaction term re-introduced multicollinearity concerns and, hence, it was removed.

In our results, we indicate effect size (i.e., odds ratio) and statistical significance of the DIF-related predictors. We flag statistical significance at three levels ($p < .05$, $p < .01$, and $p < .001$). To control for Type 1 error rate, a Bonferroni-corrected alpha for the six items on the Exponential Functions assessment is $\alpha = .008$ and a Bonferroni-corrected alpha for the seven items on the One-Variable Statistics assessment is $\alpha = .007$. For the remainder of the

manuscript, we interpreted all statistical significance for *p* values that meet the unadjusted α = .05 level to favor flagging all potential DIF items for review, but readers can utilize an adjusted α level using the information provided in the tables.

Person Fit

To evaluate person fit differences from the pre-pandemic to pandemic schooling environment, as well as the contextual school factors that may explain such differences, we first identified aberrant responses for pre-pandemic and pandemic groups of each assessment. We used the person fit statistic Zh (Drasgow, Levine, & Williams, 1985) obtained under a 2PL model framework in the mirt package (Chalmers, 2012) in R, and specifically we used the absolute value of the Zh statistic, such that increases in the person fit statistic would indicate increases in misfit. Then, we fit two multilevel regression models, one for each of the assessments. The dependent variable was the Zh person fit statistic (absolute value) and the predictors included are the same as in the multilevel DIF analysis. We fit the final multilevel regression models in Stata with maximum likelihood estimation, specified as

$$Zh_{ij} = \gamma_{00} + \gamma_{10}(PurifiedScore_{ij}) + \gamma_{20}(Pandemic_{ij}) + \gamma_{01}(Race_j) + \gamma_{02}(Title1_j) + \gamma_{03}(District\_Dummy_{j1}) + \gamma_{04}(District\_Dummy_{j2}) + u_{0j} + r_{ij.} \quad (2)$$

While the models shown in Equations 1 and 2 are quite similar, a core difference is that Equation 2 for person fit analysis has a continuous outcome, Zh_{ij}, and hence is specified in a multiple regression framework. Given Equation 2, we can explore if pandemic-related changes (Pandemic) or student knowledge on the assessment (PurifiedScore) were related to person fit, but more importantly we can explore if the school-level predictors explain differences in person fit across the pandemic groups.

Results

Differential Item Functioning

After using logistic regression DIF methods in the difR package, we identified six target items (i.e., items flagged as DIF) in the Exponential Functions assessment and seven target items in the One-Variable Statistics assessment to further evaluate for DIF in a multilevel framework while exploring contextual factors related to DIF. As many models were fit to the data (see remainder of this section), we only report for this initial single-level model whether or not the pandemic variable was statistically significant for each of the target items, the odds ratio (effect size) associated with the variable, and whether it favored the pre-pandemic or pandemic group. An item that favors the pandemic group, for example, would indicate that the item was conditionally easier for that group as compared to the pre-pandemic group. This is shown in the first row of Table 5 (for the Exponential Functions assessment) and Table 6 (for the One-Variable Statistics assessment). For the Exponential Functions assessment, all six items showed statistically significant uniform DIF in the single-level model, which aligns with the findings from the difR package. For the One-Variable Statistics assessment, Item 25 no longer showed DIF when evaluated in the Stata package, but all other item DIF findings were consistent with the difR package analysis. In the Exponential Functions assessment, five of the six items with statistically significant DIF favored the pre-pandemic group (see "-DIF" in Table 5). In the One-Variable Statistics assessment, four of the six items with statistically significant DIF favored the pre-pandemic group (see "-DIF" in Table 6).

Next, we incorporated a second level into the models to account for school clusters. The DIF results are shown in row two of Tables 5 and 6. For the Exponential Functions assessment, two of the items (Items 5 and 6) no longer showed DIF once controlling for school cluster

variable. For the One-Variable Statistics assessment, two of the items (Items 3 and 26) that showed DIF in the single-level model no longer showed DIF.

Then, we incorporated the three school-level predictors (Title 1 status, percent Black students, and district dummy variables) into level 2 of the multilevel models. Row three of Tables 5 and 6 summarize the statistical significance, effect size, and direction of these final DIF findings. For the Exponential Functions assessment, no items showed DIF once these level 2 predictors were included in the model, indicating that the DIF was explained by school-level factors in the model. For the One-Variable Statistics assessment, three items (Items 3, 10, and 16) displayed DIF even when the level 2 predictors were included in the model. Overall, eight items that showed DIF in at least one model favored the pre-pandemic group, while the remaining four items that showed DIF in at least one model favored the pandemic group.

Tables 7 and 8 show the estimated odds ratio coefficients and interclass correlation coefficients (ICCs) for all final multilevel logistic regression models with level two predictors that were fit to each target DIF item from each assessment. For the Exponential Functions assessment, we found that the percentage of Black students served by the school was a statistically significant predictor of the DIF for four of the target items (Items 1, 5, 6, and 18). Title 1 school status did not predict the DIF for any item, and only one item DIF (Item 6) was predicted by a district dummy variable. For the One-Variable Statistics assessment, we found that the percentage of Black students served by the school was a statistically significant predictor of the DIF for three of the target items (Items 3, 20, and 26). Title 1 school status predicted the DIF for one item (Item 3), and that same item's DIF was also predicted by a district dummy variable.

To ease interpretation across the myriad of DIF results, we classified each item into one of four types of DIF findings: No DIF, DIF Explained by School Clusters, DIF Explained by Contextual Factors, and Consistent DIF. Each item's classification is shown at the bottom of Tables 5 and 6. The classifications can be described as:

- No DIF: The item did not display statistically significant DIF in any of the models. One item in the One-Variable Statistics assessment displayed this type of DIF.
- DIF Explained by School Clusters: The item displayed statistically significant DIF in the single level logistic regression model, but the DIF effect disappeared once school clusters were accounted for at level 2. Two items in each of the assessments displayed this type of DIF.
- DIF Explained by Contextual Predictors: The item displayed statistically significant DIF in the single level logistic regression model and/or the multilevel model that accounted for school clusters, but the DIF effect disappeared when school-level predictors were added to level 2. Four items in the Exponential Functions assessment and two items in the One-Variable Statistics assessment displayed this type of DIF.
- Consistent DIF: The item displayed statistically significant DIF in all single level and multilevel logistic regression models. No items in the Exponential Functions assessment and two items in the One-Variable Statistics assessment displayed this type of DIF.

Person Fit

The Zh outputs used to detect aberrant responses for pre-pandemic and pandemic groups are visually displayed in Figure 1. These raw Zh values were converted to absolute values and then treated as dependent variables in two multi-level models, one for each of the assessments. Table 9 shows the multi-level model results. While the series of multilevel models used for DIF

were mimicked here when exploring person fit (albeit in a multiple regression context rather than a logistic regression context), our ultimate finding was no statistically significant differences in person fit before and during the pandemic for each assessment. In the final models, person fit was predicted at a $p < .05$ level by the purified total score in the One-Variable Statistics assessments only, and no other variables significantly predicted the person fit outcomes.

Discussion

Overall, and taking together the DIF and person fit analyses, we found for these two assessments in our project that our measurement model fits the person data well across the pre-pandemic and during pandemic instructional modes (i.e., no differences in person fit across pandemic groups), but that there are multiple items for which the model lacked invariant item parameters across such groups (i.e., multiple DIF items in results). Some of the DIF concerns across pandemic groups were explained by contextual school level factors surrounding the percentage of Black students in the schools, but other items did not show this finding and no DIF items were explained by Title 1 status of schools (i.e., our economic school-level variable). For our larger project, these findings indicate we need to reconsider using the assessment items that are not invariant across pandemic conditions, and especially when this issue has a potential differential effect on scores across schools that vary in student racial composition.

Our first research question asked if item level measurement invariance, operationalized as uniform DIF, was present across groups of students taking the assessment pre-pandemic or in the remote pandemic environment. We explored DIF in 45 test items across two classroom assessments of algebra and found that 12 (26.67%; see first row of Table 5 and Table 6) of the items displayed some form of DIF across the pre-pandemic and pandemic groups. This is not desirable, and it calls into question our ability to draw valid interpretations about the student

algebra skills when students are learning and being assessed in remote environments. However, this was not surprising as the conditions of schooling and assessment changed so dramatically.

Nine of the 12 DIF items (67%) favored the pre-pandemic groups of students (see "-DIF" in the first row of Table 5 and Table 6), which was also not surprising as the assessments were developed and evaluated for validity evidence under pre-pandemic conditions. With the drastic changes in assessment access, assessment administration, and assessment environments, coupled with likely and understandable changes in student motivation or engagement with algebra, the results of our exploration into item difficulty invariance were expected and indicated a need to revisit any planned assessment interpretations or uses. However, it should be noted that three items favored the pandemic group, so not all DIF can be interpreted under expectations that students learning during the pandemic were always disadvantaged on our assessments. Evaluating differential test functioning (DTF) is a recommended next step for assessment practitioners in a similar position, as it may be that some DIF effects accumulate or cancel out at the total test score level (Drasgow, 1987; Rosnowski, 1987). Item removal for scoring may also be considered based on such DIF and DTF analysis results. In addition, for projects that have access to the assessment content and student examinees (which was not the case in this study), conducting content analyses and response process analyses may help to understand the reasons for each DIF finding.

As for contextual aspects of DIF (research question 3), we found that the percentage of Black students served within a school was a significant predictor of conditional item performance for seven of the twelve DIF items (58.33%; see "Percent Black" row in Table 7 and Table 8), and that the prediction fully explained the DIF effect for 3 of the twelve DIF items (25%) meaning that once this predictor was included in the model the DIF effect disappeared

(see items 1 and 18 of the Exponential Functions assessment and item 20 of the One-Variable Statistics assessment, where the DIF effect becomes non-statistically significant once the Percent Black variable is added to the model). This is an important finding because it indicates that some, but certainly not all, of our measurement invariance issues are related to conditional impacts on item response across schools serving different racial compositions of students (i.e., schools serving students with differential percentages of Black students have differential item performance expectations even when considering only students who have the same purified test score). While further exploration into this phenomenon may be able to further explain why this was occurring, it is important to recognize that the multicollinearity of so many of our school contextual variables means that our Percent Black school variable is capturing more than just student racial composition. Taking this only as a descriptive, not causal finding, our research team on our larger project likely will remove these items for future scoring regardless of the "why" behind the findings, as both the DIF and the conditional impact of school-level racial variables are cause enough for concern of valid and consistent score interpretation across schools.

Overall, given the DIF results of the study, our findings indicate that our item assessment data contain conditional item difficulty parameter differences that go beyond differences in instructional mode to include differences in school-level student race. For this reason, we consider the DIF findings to likely indicate a form of item bias. As the goal is to measure algebra knowledge, knowing that there is likely item bias across racial lines calls into question not only the validity and fairness of any of our inferences from test scores including such items, but also requires a question surrounding ethics in our research that hinges on our measurements of student algebra knowledge. If we are not measuring the same thing across different instructional

modes of learning and assessment, and student race is sometimes an additional explanation of such measurement problems, might some of our project's research findings be explained by measurement issues across pandemic and/or racial lines? At a minimum, it is critical that we conduct these analyses and remove all items with such issues from the assessment scoring. But from a broader perspective, it is critical that we continue to investigate and share instances in which conditional assessment performance is correlated to race, income, or other variables capturing school context variables that we aim to measure invariantly across (Garcia & Weiss, 2020).

Our second research question focused on another potential concern for assessment introduced by the health pandemic, that of person fit within model-based classroom assessment frameworks. While we located individuals with some person fit concerns in our assessments, they were not explained by differences in pre-pandemic and pandemic assessment conditions and they were not predicted by the contextual school-level factors we were able to explore in our study (research question 3).

A limitation of this study lies in our access to data on variables of interest. Ideally, student-level demographics would be incorporated into our models, allowing for additional research questions at that level, but this was unfortunately not permitted in the IRB for our larger project. Also, all school-level variables were obtained through public information, whereas access to additional variables could have been beneficial for multiple reasons, including the possibility of being able to explore intersectionality of these school-level variables and the possibility of using school-level variables that tap into more constructs without introducing multicollinearity. Finally, while covariate control alone would not permit us to draw causal inferences from our results, having additional variables could have ruled out additional

alternative explanations for some of the relationships we found between school-contextual factors and psychometric outcomes. For this reason and more, it is important that readers understand that the predictions we found between percentage of Black students served by a school and conditional item performance can indicate that this item issue is sometimes related to this variable, but the cause of such relationship is unknown and may be explained by non-racial factors.

Another limitation to our study was the exclusion of students who did not complete the full assessment, and the fact that we had no student-level variables to better understand the types of students who did not complete the full assessment. The logfile data in the larger project is unstructured, large, ongoing, and not entirely in control of researchers, as is common in the "digital ocean" of educational data (Cope & Kalantzis, 2016; DiCerbo & Behrens, 2014; National Forum on Educational Statistics, 2015). This results in large amounts of missingness that we have evidence is related to various systematic factors in the data, making it nonignorable (Xue et al., 2022). However, the methodological literature on evaluating issues of measurement invariance and person fit in unstructured data is in its infancy, as is much of the computational psychometrics literature (von Davier et al., 2019). Similarly, traditional multiple imputation methods were not directly applicable in our data set for a variety of methodological challenges. Hence, our findings only generalize to the students in our project who were engaged in the assessments enough to complete them and/or had teachers who provided motivations to complete assessments. Having said that, one could always argue that some data were not included in a study that uses logfile data given the ongoing nature of the logfiles for which any one-time data extraction process will result in a time-bounded cross-section of that ongoing data collection.

A delimitation of our study is that all data came from an algebra learning tool and two embedded algebra assessments. It is unknown how our answers to the research questions would have changed if the constructs of measurement were unrelated to algebra. Certainly, we do not assume that our results generalize to other assessments in general, and especially if the assessments are not related to algebra, upper middle school or lower high school students, and virtual learning environments. Rather than generalize our results in any way, our aim to inform other practitioners of assessment is focused on encouraging such persons to explore their data with similar methods, but with no expectation that substantive findings will align with ours. However, it would be of interest to the field to know if it is a common occurrence that assessment items lacked invariance across heterogenous instructional modes and if such invariance is predicted by school contextual variables.

Scholarly Significance

The global COVID-19 health pandemic introduced remote learning and assessment at-scale in K-12 education systems, and it is conjectured that while the fully remote education shift from 2020 was only temporary, the school systems will not fully return to exactly what they were prior to 2020 for the foreseeable future (U. S. Department of Education, Office of Civil Rights, 2021). Combined with the spotlight on inequalities in the US education system, many questions are being raised about educational assessment across different modes of instruction and assessment, and in particular how such changes may perpetuate problems of educational inequalities (García & Weiss, 2020; U. S. Department of Education, Office of Civil Rights, 2021). For those using classroom assessment data to inform instruction or research, questions of valid interpretations of assessment scores in these new educational contexts have been paramount. Can we assume that students approached the assessments in the same way as they did

pre-pandemic? Conversely, can we assume that assessment features have remained consistent across the changing education and testing environments, and across different groups of students learning in such environments? In this study, we explored three research questions to help us address these broad questions in our research project and to share our findings with others.

With the drastic changes in assessment access, assessment administration, and assessment environments, coupled with likely and understandable changes in student motivation or engagement with algebra, the results of our exploration into item difficulty invariance were expected and indicate a need to revisit any planned assessment interpretations or uses. School shutdowns and remote learning caused by the pandemic resulted in scholars, educators, and administrators operating in unchartered territory. Hence, we plan to continue our work in exploring and revisiting psychometric challenges as they relate to contextual features. We recommend that educational measurement researchers develop and evaluate methods for exploring all psychometric issues related to measurement invariance, measurement model-data fit, and more that can negatively impact the validity of interpretations from classroom assessment scores. We particularly recommend that these measurement problems be explored in the light of educational inequality, to ensure that the field and practice of assessment continues to seek valid measurement of all students. This work is much needed in an educational environment that has quickly adopted educational technology platforms to allow for remote learning and assessment, but that still must assess students in equitable ways to validly and fairly inform research and instruction.

References

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick's (Eds.), *Statistical Theories of Mental Test Scores* (397-479). Reading, Mass, Addison-Wesley.

Brennan, R. L. (Ed.). (2006). *Educational Measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.

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

Chen, M. Y., & Zumbo, B. D. (2017). Ecological framework of item responding as validity evidence: An application of multilevel DIF modeling using PISA data. In B. D. Zumbo & A. M. Hubley's (Eds.), *Understanding and investigating response processes in validation research* (53–68). New York, NY: Springer.

Clauser, B., & Mazor, K. M. (2005). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice, 17,* 31-44.

Cope, B., & Kalantzis, M. (2016). Big Data comes to school: Implications for learning, assessment, and research. *AERA Open, 2,* 1-19.

DiCerbo, K. E., & Behrens, J. T. (2014). *Impacts of the digital ocean on education.* London, UK: Pearson.

Dixon-Román, E. (2020). A haunting logic psychometrics: Toward the speculative and indeterminacy of blackness in measurement. *Educational Measurement: Issues and Practice, 39*(3), 94-96.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19-29.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

García, E., & Weiss, E. (2020). *COVID-19 and student performance, equity, and U.S. education policy: Lessons from pre-pandemic research to inform relief, recovery, and rebuilding* (Report No. 205622). Economic Policy Institute. https://www.epi.org/publication/the-

consequences-of-the-covid-19-pandemic-for-education-performance-and-equity-in-the-united-states-what-can-we-learn-from-pre-pandemic-research-to-inform-relief-recovery-and-rebuilding/

Goldstein H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. Journal of Educational Measurement, 20, 369-377.

Harris, D. (Ed.). (2020). Special section: Impact of COVID19 on educational measurement [Special issue]. *Educational Measurement: Issues and Practice, 39*(3).

Huggins-Manley, A. C., Beal, C. R., D'Mello, S. K., Leite, W. L., Cetin-Berber, D. D., Kim, D., & McNamara, D. S. (2019). A commentary on construct validity when using operational virtual learning environment data in effectiveness studies. *Journal of Research on Educational Effectiveness, 12,* 750-759.

Karabatsos, G. (2010). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16,* 277-298.

Lastinger Center for Learning, & University of Florida. (2019). *Algebra Nation*. Retrieved 9/20/2019 from http://lastingercenter.com/portfolio/algebra-nation-2/

Magis, D., Beland, S., & Raiche, G. (2020). difR (R package). https://cran.r-project.org/web/packages/difR/difR.pdf.

National Forum on Education Statistics. (2015). *Forum Guide to Elementary/Secondary Virtual Education Data*. (NFES 2016-095). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Paek, I., & Guo, H. (2011). Accuracy of DIF estimates and power in unbalanced designs using the mantel-haenszel DIF detection procedure. *Applied Psychological Measurement, 35,* 518-535.

Penfield, R. D., & Camilli, G. (2006). *Differential item functioning and item bias.* In C. R. Rao & S. Sinharay's (Eds.) *Handbook of Statistics*, Vol. 6 (125-167). Amsterdam, The Netherlands: Elsevier.

R Development Core Team (2020). *R: A language and environment for statistical computing, reference index version 4.1.0.* R Foundation for Statistical Computing. Vienna: Austria. URL http://www.R-project.org.

Randall, J. (2021). "Color-Neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical anti-racist lens. *Educational Measurement: Issues and Practice* (online first). https://doi.org/10.1111/emip.12429

Rosnowski, M. (1987). Use of tests manifesting sex differences as measures of intelligence: Implications for measurement bias. *Journal of Applied Psychology, 72,* 480-483.

Sireci, S. (2021). NCME Presidential Address 2020: Valuing educational measurement. *Educational Measurement: Issues and Practice, 40*(1), 7-16.

StataCorp. (2021). *Stata Statistical Software: Release 17*. College Station, TX: StataCorp LLC.

U. S. Department of Education. (2020a). (Notice of waiver approval for Florida Department of Education). Author. Retrieved May 26, 2021 from http://www.fldoe.org/core/fileparse.php/19861/urlt/FLFiscalWaiverResponse.pdf.

U. S. Department of Education. (2020b). *Notice of waivers granted under Section 8401 of the Elementary and Secondary Education Act of 1965 (ESEA)* (Notice 85 FR 29935; Document No. 2020-10740). Author. Retrieved May 26, 2021 from https://www.federalregister.gov/documents/2020/05/19/2020-10740/notice-of-waivers-granted-under-section-8401-of-the-elementary-and-secondary-education-act-of-1965.

U. S. Department of Education, Office for Civil Rights, *Education in a Pandemic: The Disparate Impacts of COVID-19 on America's Students ,* 86 Fed.Reg. 7009, 7009 (June 9, 2021), https://www2.ed.gov/about/offices/list/ocr/docs/20210608-impacts-of-covid19.pdf.

Vo, T. T., & French, B. F. (2021). An ecological framework for item responding within the context of a youth risk and needs assessment. *Educational Measurement: Issues and Practice* (online first). https://doi.org/10.1111/emip.12426

von Davier, A. A., Deonovic, B., Yudelson, M., Polyak, S. T., Woo, A. (2019). Computational psychometric approach to holistic learning and assessment systems. *Frontiers in Education*, *4*. https://doi.org/10.3389/feduc.2019.00069.

Xue, K., Huggins-Manley, A. C., & Leite, W. L. (2022). Semi-supervised learning method to adjust biased item difficulty estimates caused by nonignorable missingness in a virtual learning environment. *Educational and Psychological Measurement, 82,* 539-567. https://doi.org/10.1177/00131644211020494

Zumbo, B. D. *(*2007*).* Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223–233.

Table 1: Item Means and Correlations to Total Test Scores

Exponential Functions Assessment			One-Variable Statistics Assessment		
Item	Mean	Correlation	Item	Mean	Correlation
1	.43	.58	1	.58	.50
2	.53	.53	2	.26	.43
3	.73	.37	3	.18	.40
4	.33	.59	4	.25	.42
5	.32	.61	5	.20	.43
6	.29	.62	6	.39	.52
7	.42	.40	7	.73	.55
8	.69	.52	8	.33	.47
9	.31	.51	9	.45	.57
10	.33	.53	10	.20	.39
11	.41	.56	11	.38	.54
12	.24	.50	12	.75	.55
13	.49	.60	13	.37	.60
14	.48	.53	14	.20	.46
15	.58	.41	15	.27	.47
16	.50	.59	16	.37	.61
17	.65	.45	17	.57	.54
18	.36	.60	18	.47	.65
---	---	---	19	.35	.58
---	---	---	20	.44	.59
---	---	---	21	.42	.63
---	---	---	22	.36	.63
---	---	---	23	.25	.56
---	---	---	24	.38	.58
---	---	---	25	.49	.61
---	---	---	26	.11	.35
---	---	---	27	.38	.64

Table 2: Descriptive Statistics for School-level Predictors

Variable	Exponential Functions Assessment					One-Variable Statistics Assessment				
	Obs	Mean	SD	Min	Max	Obs	Mean	SD	Min	Max
Pandemic	927	.28	.45	0	1	1,019	.72	.45	0	1
District1	262	.60	.49	0	1	223	1	0	1	1
District2	207	.46	.50	0	1	191	.34	.47	0	1
District3	458	.01	.08	0	1	605	.74	.44	0	1
Percent Hispanic	927	.28	.16	.05	.73	1,019	.28	.13	.05	.73
District1	262	.36	.13	.05	.6	223	.37	.13	.07	.6
District2	207	.13	.06	.05	.21	191	.15	.06	.05	.21
District3	458	.30	.15	.15	.73	605	.28	.11	.15	.73
Percent Black	927	.33	.25	.03	.93	1,019	.28	.25	.04	.89
District1	262	.37	.19	.04	.93	223	.31	.26	.04	.89
District2	207	.47	.29	.09	.83	191	.35	.27	.09	.83
District3	458	.25	.23	.03	.72	605	.25	.23	.05	.72
Percent White	927	.30	.19	0	.67	1,019	.36	.21	.02	.67
District1	262	.18	.10	0	.31	223	.23	.18	.02	.50
District2	207	.32	.22	.08	.67	191	.41	.22	.08	.67
District3	458	.36	.19	.06	.65	605	.38	.20	.06	.65
Title 1	927	.50	.50	0	1	1,019	.44	.50	0	1
District1	262	.66	.48	0	1	223	.66	.48	0	1
District2	207	.73	.44	0	1	191	.66	.47	0	1
District3	458	.30	.46	0	1	605	.29	.45	0	1
Ecodis	927	.54	.24	.16	1	1,019	.49	.23	.16	1
District1	262	.54	.16	.16	.89	223	.47	.19	.16	.83
District2	207	.71	.26	.32	1	191	.62	.25	.32	1
District3	458	.46	.23	.24	.87	605	.45	.23	.24	.89
ELL	927	.06	.06	0	.21	1,019	.06	.06	0	.22
District1	262	.05	.02	.02	.12	223	.04	.01	.03	.06
District2	207	.04	.03	0	.13	191	.04	.03	0	.09
District3	458	.09	.08	0	.21	605	.07	.07	0	.22
Disability	927	.13	.05	.04	.24	1,019	.14	.06	.04	.26
District1	262	.14	.03	.07	.18	223	.14	.02	.07	.16
District2	207	.15	.02	.13	.19	191	.17	.04	.13	.26
District3	458	.12	.06	.04	.24	605	.14	.07	.04	.26
Purified score	927	5.78	3.02	0	12	1,019	8.04	5.16	0	20
District1	262	5.08	3.12	0	12	223	7.12	5.34	0	20
District2	207	5.27	3.15	0	12	191	8.12	4.97	0	20
District3	458	6.40	2.76	0	12	605	8.36	5.13	0	20
Total score	927	8.10	4.55	0	18	1,019	10.10	6.60	0	27
District1	262	6.98	4.55	0	18	223	8.94	6.88	0	26
District2	207	7.25	4.70	0	18	191	10.18	6.47	0	27
District3	458	9.12	4.25	0	18	605	10.50	6.50	0	26

Table 3: Pairwise Correlation Matrix of School-level Predictors

	Exponential Functions							One-Variable Statistics						
	Title1	Eco-dis	ELL	Dis-ability	His-panic	Black	White	Title1	Eco-dis	ELL	Di-sability	His-panic	Black	White
Title1	1.00							1.00						
Ecodis	.74	1.00						.87	1.00					
ELL	.35	.42	1.00					.49	.58	1.00				
Disability	.55	.60	.56	1.00				.49	.64	.56	1.00			
Hispanic	-.06	-.17	.38	.08	1.00			-.04	-.18	.28	.08	1.00		
Black	.64	.81	.16	.32	-.53	1.00		.80	.86	.37	.36	-.41	1.00	
White	-.71	-.76	-.38	-.33	-.12	-.76	1.00	-.85	-.80	-.52	-.36	-.14	-.83	1.00

Table 4: Multicollinearity VIF Check using School-level OLS Regression

	Exponential Functions	One-Variable Statistics
Title 1		
Ecodis		Y
ELL		
Disability		
Percent Hispanic	Y	Y
Percent Black	Y	Y
Percent White	Y	Y

Note. Y represents VIF > 10.

Table 5: DIF Classification for Six Target Items on the Exponential Functions Assessment

Model	Exponential Functions Assessment					
	Item 1	Item 5	Item 6	Item 7	Item 13	Item 18
Single-level Model	-DIF**	-DIF*	-DIF**	+DIF**	-DIF**	-DIF***
Pandemic Odds ratio (se)	.58(.10)	.65(.13)	.51(.11)	1.64(.26)	.54(.10)	.49(.1)
Multi-level model, no school-level predictors	-DIF*	--	--	+DIF**	-DIF*	-DIF*
Pandemic Odds ratio (se)	0.61(.14)	.71(.17)	.60(.17)	1.64(.26)	.57(.12)	.48(.15)
[ICC]	[.04]	[.02]	[.05]	[.00]	[.02]	[.09]
Multi-level model	--	--	--	--	--	--
Pandemic Odds ratio (se)	.69(.18)	1.03(.28)	.73(.22)	1.50(.32)	.79(.20)	.63(.21)
[ICC]	[.01]	[.00]	[.01]	[.00]	[.00]	[.05]
Classification for Interpretation	DIF Explained by Contextual Predictors	DIF Explained by School Clusters	DIF Explained by School Clusters	DIF Explained by Contextual Predictors	DIF Explained by Contextual Predictors	DIF Explained by Contextual Predictors

* $p < .05$; ** $p < .01$; *** $p < .001$ (alpha under Bonferonni correction is .008, requiring significance of ***)
Note: negative DIF favored the pre-pandemic group and positive DIF favored the pandemic group.

Table 6: DIF Classification for Seven Target Items on the One-Variable Statistics Assessment

	One-Variable Statistics						
	Item 3	Item 10	Item 15	Item 16	Item 20	Item 25	Item 26
Single-level Model Pandemic Odds ratio (se)	-DIF* .65(.12)	-DIF** .54(.10)	+DIF* 1.59(.28)	-DIF** .59(.10)	+DIF*** 1.85(.32)	-- 1.38(.24)	-DIF*** .40(.09)
Multi-level model, no school-level predictors Pandemic Odds ratio (se) [ICC]	-- .74(.20) [.07]	-DIF** .54(.10) [.00]	+DIF* 1.59(.31) [.01]	-DIF** .59(.10) [.00]	+DIF** 1.76(.36) [.01]	-- 1.24(.29) [.05]	-- .49(.22) [.24]
Multi-level model Pandemic Odds ratio (se) [ICC]	-DIF* .57(.14) [.00]	-DIF* .64(.14) [.00]	-- 1.36(.30) [.00]	-DIF** .56(.12) [.00]	-- 1.50(.31) [.00]	-- 1.13(.26) [.01]	-- .51(.23) [.17]
Classification for Interpretation	DIF Explained by School Clusters (partially)	Consistent DIF	DIF Explained by Contextual Predictors	Consistent DIF	DIF Explained by Contextual Predictors	No DIF	DIF Explained by School Clusters

* $p < .05$; ** $p < .01$; *** $p < .001$ (alpha under Bonferonni correction is .007, requiring significance of ***)
Note: negative DIF favored the pre-pandemic group and positive DIF favored the pandemic group.

Table 7: Multilevel Logistic Regression Model Results for Six Target Items on the Exponential Functions Assessment

Predictor/ ICC	Odds Ratio (Standard Error)/ ICC (Standard Error)					
	Item 1	Item 5	Item 6	Item 7	Item 13	Item 18
PurifiedScore	1.44 (.05)***	1.51 (.05)***	1.54 (.06)***	1.27 (.03)***	1.51 (.05)***	1.51 (.05)***
Pandemic	.69 (.18)	1.03 (.28)	.73 (.22)	1.50 (.32)	.79 (.20)	.63 (.21)
Title 1	1.45 (.35)	1.02 (.25)	.66 (.18)	.73 (.14)	.81 (.18)	1.08 (.38)
Percent Black	.19 (.10)***	.30 (.16)*	.24 (.14)*	1.44 (.58)	.68 (.33)	.19 (.14)*
District 1	1.04 (.27)	.64 (.18)	1.29 (.38)	1.25 (.26)	.76 (.18)	1.00 (.35)
District 2	.75 (.19)	1.24 (.30)	1.85 (.52)*	1.35 (.28)	.72 (.17)	.77 (.27)
ICC	.01 (.01)	.00 (.00)	.01 (.01)	.00 (.00)	.00 (.01)	.05 (.04)

* $p < .05$; ** $p < .01$; *** $p < .001$(alpha under Bonferonni correction is .008, requiring significance of ***)
Note: negative DIF favored the pre-pandemic group and positive DIF favored the pandemic group.

Table 8: Multilevel Logistic Regression Model Results for Seven Target Items on the One-Variable Statistics Assessment

Predictor/ ICC	Odds Ratio (Standard Error)/ ICC (Standard Error)						
	Item 3	Item 10	Item 15	Item 16	Item 20	Item 25	Item 26
PurifiedScore	1.23 (.02)***	1.19 (.02)***	1.22 (.02)***	1.34 (.03)***	1.30 (.02)***	1.31 (.02)***	1.25 (.03)***
Pandemic	.57 (.14)*	.64 (.14)*	1.36 (.30)	.56 (.12)**	1.50 (.31)	1.13 (.26)	.51 (.23)
Title 1	.34 (.14)**	1.07 (.35)	.93 (.30)	1.28 (.40)	1.44 (.42)	.67 (.22)	.53 (.44)
Percent Black	21.84 (15.37)***	.85 (.54)	1.09 (.70)	.34 (.22)	.20 (.12)**	.47 (.30)	15.28 (20.95)*
District 1	2.01 (.56)*	.93 (.24)	1.21 (.29)	.98 (.24)	.76 (.17)	.96 (.24)	1.37 (.89)
District 2	.70 (.21)	1.45 (.34)	.76 (.20)	1.10 (.26)	.70 (.16)	1.05 (.28)	1.10 (.67)
ICC	.00 (.01)	.00 (.00)	.00 (.01)	.00 (.00)	.00 (.00)	.01 (.01)	.17 (.08)

* $p < .05$; ** $p < .01$; *** $p < .001$(alpha under Bonferonni correction is .008, requiring significance of ***)
Note: negative DIF favored the pre-pandemic group and positive DIF favored the pandemic group.

Table 9: Person Fit Multilevel Model Results for Each Assessment

	Exponential Functions		One-Variable Statistics	
	Coefficient	Standard Error	Coefficient	Standard Error
PurifiedScore	-.00	.01	.02***	.00
Pandemic	-.10	.05	.01	.05
Title 1	-.06	.05	.04	.06
Percent Black	.13	.10	.06	.12
District 1	.06	.05	-.01	.05
District 2	.08	.05	-.00	.05
ICC (no sig.)	.00	.01	.00	.00

* $p < .05$; ** $p < .01$; *** $p < .001$

Figure 1

Zh Statistic for Aberrant Responses