Examining School Accountability: Discriminating Schools' A-F Report Card Grades

Timothy D. Folger
Audrey Conway Roberts
Jonathan D. Bostic
Bowling Green State University

***T**he COVID-19 pandemic disrupted many school accountability systems that rely on student-level achievement data. Many states encountered uncertainty about how to meet federal accountability requirements without typical school data. Prior research provides evidence that student achievement is correlated to students' social background, which raises concerns about the predictive bias of accountability systems. This mixed-methods study (a) examines the predictive ability of non-achievement-based variables (i.e., students' social background) on school districts' report card letter grade in Ohio, and (b) explores educators' perceptions of report card grades. Results suggest that social background and community demographic variables have a significant impact on measures of school accountability.*

Introduction

Accountability testing became widespread in the United States after the signing of the No Child Left Behind Act of 2001 (NCLB) and has continued throughout subsequent reauthorizations. Current federal mandates allow states the flexibility of developing their own educational accountability system but require states to include student achievement data from large-scale standardized tests to measure school and district performance (Every Student Succeed Act [ESSA], 2015). The COVID-19 pandemic, however, disrupted the achievement-based accountability systems that were in place to some degree. "States face considerable uncertainty about how to meet federal and state accountability requirements for [the 2020-2021] school year and beyond" (Lake & Worthen, 2021, p. l). Often, the clearest way states communicate the performance of a public school district is by an annual report card, which assigns A-F letter grades as indicators of school and district performance (Murray & Howe, 2017). As the typical cycle of standardized testing has been disrupted, this study investigates the predictive ability of non-achievement-based variables (e.g., enrollment and median household income) on composite grades of school performance. The rationale for this study is two-fold. First, a statistical model based on social and community demographic (i.e., non-achievement) variables that can reliably - and accurately - predict measures of school accountability contributes evidence of the potential inequities permeating education systems. Thus, this research takes a strong access and equity perspective (National Council of Teachers of Mathematics [NCTM], 2014). Second, exploring the predictive ability of non-achievement-based variables encourages dialogue about whether current systems of accountability are valid measures of school performance and quality.

School Accountability Report Cards

Sixteen states, including Ohio whose data were used in the present study, have used A-F letter grades, or a similar system, to evaluate school performance (Adams et al., 2016; Murray & Howe, 2017). Ohio Revised Code (3302.03) defines the composite letter grade as the overall performance of a school or district (Ohio Revised Code, 2018). Common rationales arise from states implementing A-F systems of accountability. An easy-to-understand system of accountability empowers parents to make educational decisions based on the overall academic performance of schools, and states implementing A-F systems of accountability argue that the system helps identify ways to improve the quality of education being provided by school districts (e.g., Arizona Department of Education, 2021; Murray & Howe, 2017; Utah State Board of Education, 2020). Table 1 documents the purposes of school and district report cards identified by the Ohio Department of Education (ODE). To summarize, the purpose of the A-F accountability system is to measure school performance and the quality of education being provided to students.

Murray and Howe (2017) synthesized the arguments in favor of A-F systems of accountability. They concluded, "The chorus in favor of A-F systems seems to be singing the same refrain: A-F systems are supposedly clear, concise systems that let everyone know how schools are doing and encourage parents to be involved in school choices and systems" (2017, p. 6). Adams and colleagues (2016) compared achievement difference between letter grades while controlling for variance in school composition. For example, a comparison of achievement differences among students receiving free or reduced lunch (FRL) indicated larger achievement gaps in higher rated school districts. More notably, FRL students in the lowest rated schools outperformed FRL students in the highest rated schools (Adams et al., 2016).

Table 1
Purposes of the grade card accountability system (Ohio Department of Education, 2021)

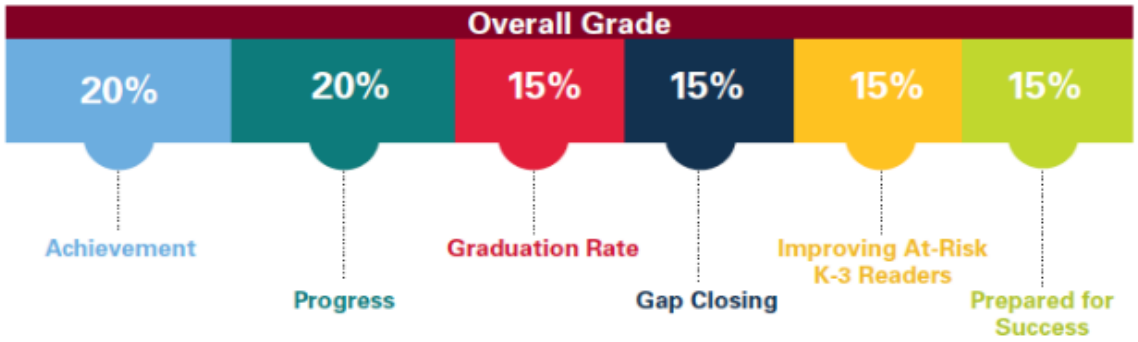
Purpose	Description
Student growth and achievement	To provide communities a picture of school and district progress in raising student achievement and preparing students for the future.
Identify strengths and weaknesses	To provide educators, school administrators, and families information about the strengths and weaknesses of school performance.
Quality of education	To provide parents and schools an understanding about the quality of education being provided to students.

Ohio Revised Code (3302.03) mandates an overall grade to be calculated from six components of school performance for all public districts and individual schools. Student achievement data from end-of-grade (EOG) and end-of-course (EOC) tests hold significant weight in the calculation of school districts' final composite grades. Figure 1 displays the six components that are used to determine final grades and the weight of each component in calculating the final composite grade. The final letter grade, as well as all six component grades, earn a score between 0 and 5 points. For each component grade, the points earned are classified by the percentage

score on the respective measure's grading scale. A comprehensive review of how each component grade is calculated is beyond the scope of this review but is described in a technical document released by the Ohio Department of Education (ODE, 2020). The resulting score indicates the letter grade earned (see Table 2). The points contributing to the final letter grade are calculated by multiplying the weights of each component, shown in figure 1, by the points earned for each corresponding component. A final composite grade is determined by calculating the sum of the six components' weighted points. See Table 3 for an example of calculating a final letter grade. Student achievement, based on EOG and EOC standardized test results, is its own component of the report card. However, interpretations of EOG and EOC test results are also used to measure other constructs such as "gap closing" and "progress". *Gap closing* measures the improvements in achievement by subgroups of students (ODE, 2021).

Figure 1

Six components of composite grades (ODE, 2021)



Progress, or value-added, measures student growth over the course of the year. More specifically, value-added scores measure students' academic growth based on achievement at two points in time (ODE, 2020). Proponents of value-added models argue this allows students to serve as their own 'control' to account for extraneous variables (i.e., students' social backgrounds) contributing to academic achievement (Sanders & Horn, 1994, 1998). Ballou and colleagues (2004) determined "controlling for SES and demographic factors at the student level makes very little difference to teacher effects estimated by the Tennessee Value-Added Assessment System (TVAAS)" (p. 60). That is, adjusting the TVAAS model to statistically control for student SES and demographic variables did not result in significantly differently value-added scores for teachers (Ballou et al., 2004). Scholars who challenge the merits of value-added models raise concerns regarding validity and reliability, or consistency of scores over time, finding "teachers classified as 'effective' one year might have a 25% to 59% chance of being classified as "ineffective" the next year, or vice versa" (Amrein-Beardsley & Close, 2019, p. 872). Such variation suggests value-added scores may not be reliable indicators of teacher performance. Additionally, critics claim value-added models rely on "heroic" assumptions when making inferences about teachers' direct impacts on student achievement over time (Amrein-Beardsley & Close, 2019; Rubin et al., 2004). In determining schools' report card letter grade, it is clear how interpretations about student achievement from standardized test results are used in a multitude of ways and are the dominant feature in this accountability model.

Table 2
Letter grade designations based on total points (ODE, 2020)

Total Points	Letter Grade
4.125 – 5.000	A
3.125 – 4.124	B
2.125 – 3.124	C
1.125 – 2.124	D
0 – 1.124	F

Table 3
Calculating a school district's final letter grade (ODE, 2020)

Component	Points Earned	Weight	Weighted Points
Achievement	2.625 (C)	0.20	0.5250
Progress	3.400 (B)	0.20	0.6800
Graduation	4.000 (B)	0.15	0.6000
Gap Closing	3.250 (B)	0.15	0.4875
Improving At-Risk K-3 Readers	3.000 (C)	0.15	0.4500
Prepared for Success	3.250 (B)	0.15	0.4875
Total Weighted Points			3.230 (B)

Student Achievement

Differences in academic achievement strongly correlate with students' social background (Broer et al., 2019; Caldas & Bankston, 1997; May, 2006). Socio-economic status (SES), "the social standing or class of an individual or group" (American Psychological Association [APA], 2021), has been measured from a variety of sources including median household income, parents' educational attainment and professional occupation, and even home possessions (Broer et al., 2019). May (2006) examined the relationship between SES and fourth grade reading achievement using SES indicators such as the percentage of students identified as economically disadvantaged and median household income. Broer and colleagues (2019) analyzed the relationship between SES and student achievement using twenty years of data from the Trends in International Mathematics and Science Study (TIMSS). The results from such research studies

are clear, the correlation between SES and academic achievement is strong and statistically significant (Broer et al., 2019; May, 2006). In addition to evidence of the correlation between academic achievement and SES, attending school with classmates from high SES may positively affect academic achievement, regardless of one's own social background (Caldas & Bankston, 1997). Caldas and Bankston (1997) found that all students displayed greater academic achievement when attending a high SES school district. More specifically, "The effect of schoolmates' family social status on achievement is significant and substantial, and only slightly smaller than an individual's own family background status" (Caldas & Bankston, 1997, p. 275).

Gaps in achievement based on SES continue to exist in lieu of initiatives meant to close such gaps. Analysis of TIMSS data indicated the mathematics achievement gap has neither increased nor decreased from 1995 to 2015 (Broer et al., 2019; Chmielewski, 2019). Gaps in achievement related to students' social background raise concern about how accountability systems measure school performance. Wiliam (2010) states, "when a person, organization, or entity is accountable, they can be expected or required to render an account of their actions. The two immediate questions that follow are 'to whom?' and 'for what?'" (p. 108). It stands to reason that report card grades should reflect aspects of education controlled by schools and educators (i.e., teachers). Report card grades should not simply reward school districts serving socially privileged students (Adams et al., 2016). However, the degree to which report card grades yield accurate depictions of school quality and performance is a question regarding the validity of report card grades.

Valid Interpretations and Uses of Test Scores

Validity is defined as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association [AERA] et al., 2014, p. 11). Validity is an attribute of the proposed interpretation and use of test scores, not the test itself (AERA et al., 2014). The A-F report card system uses interpretations of achievement test results to draw inferences about school performance and quality. As such, the validity of the A-F report card system warrants investigation. *The Standards for Educational and Psychological Testing [The Standards]* outline five sources of validity evidence: test content, response processes, internal structure, relations to other variables, and consequential/bias (AERA et al., 2014). The validity of a proposed interpretation and use of test results depends on the quantity and quality of evidence supporting the proposed interpretation and use (AERA et al., 2014; Lavery et al., 2019; Kane, 2013, 2020). For instance, state-level achievement tests may have sufficient validity evidence to measure students' knowledge of grade-level content standards, but insufficient evidence to support the use of test scores as indicators of school effectiveness (Adams et al., 2016). Consequently, complex interpretations and uses of test scores require a more robust validity argument (AERA et al., 2014; Lavery et al., 2019; Kane, 2013). Lavery and colleagues (2019) state,

> assessments that inform high-stakes decisions, or decisions for which the consequences of 'getting it wrong' are either grave or costly, require that substantially more time, attention, and resources be devoted to ensuring they produce accurate scores that are valid for the intended interpretations and uses. (p. 13)

Achievement tests tend to be used for high-stakes purposes while classroom assessment tend to have low-stakes attached to them. For example, a classroom assessment measuring students' ability to solve algebraic equations requires relatively simple generalizations. Primarily, that the scores from the observed performance can be extrapolated to the target domain (Kane, 2013). Using classroom assessment results to plan future instruction and/or intervention has relatively low stakes. The cost of wrongly deciding students are prepared for more advanced instruction will likely lead to re-teaching and/or providing intervention in the future. In contrast, extrapolating student-level achievement data to make causal inferences about the quality of education being provided by a school district requires complex generalizations because the unit of analysis is not the test-taker (Kane, 2013). Achievement tests are high stakes when results are used to determine funding, make employee decisions, or used by parents to make decisions about their child's education. Even test results reported in aggregate meant to be used for low-stakes purposes, such as measures of school performance, may inadvertently produce high-stakes uses of test scores by informing judgments about personnel and/or program quality (AERA et al., 2014). Consequently, such judgments are likely to shape policy decisions. The stakes of the test are closely related to both the intended and unintended consequences of how test scores are used (AERA et al., 2014).

The Ohio Department of Education publishes an annual technical report which states the intended uses of achievement test scores. Intended uses for state test scores include school accountability, feedback about student and class performance, and evaluation of teacher performance (Cambium Assessment, 2020). Test developers are expected to evaluate the claims inherent to, and validity evidence for and against, each intended use of test scores (AERA et al., 2014; Kane, 2013). As identified in the annual technical report, school accountability systems, such as the A-F report card, use interpretations of test results to draw inferences about school performance. The logic behind accountability testing is quite simple, that students receiving higher quality education will display higher achievement. However, accountability testing requires the converse to hold true - that greater student achievement is indicative of higher quality education (Wiliam, 2010). The *Standards* use the term *predictive bias* to describe differences “in the patterns of associations between test scores and other variables for different groups, bringing with it concerns about bias in the inferences drawn from the use of test scores” (AERA et al., 2014, p. 51). Therefore, patterns in student achievement based on characteristics such as socioeconomic status merit close examination when considering the validity of using achievement test results to make judgments about school performance and quality.

This study examines patterns of associations between composite grades of school district performance and non-achievement-based variables (e.g., SES and race/ethnicity). This study is not a systematic review of all validity evidence for and against the report card system as an interpretation and use of achievement test results. We contribute to the "logical debate" that is validation (Cronbach, 1988) by specifically reflecting on one source of validity evidence: relationships to other variables. Evidence based on relations to other variables is/are important when "the intended interpretation for a given use implies that the construct should [or should not] be related to some other variable" (AERA et al., 2014, p. 16). For instance, evidence based on relations to other variables may support the claim that "test scores are not unduly influenced by ancillary variables" (e.g., socioeconomic status; AERA et al., 2014, p. 12). Such patterns of association are not intended to imply that a testing program itself is biased or unfair (AERA et

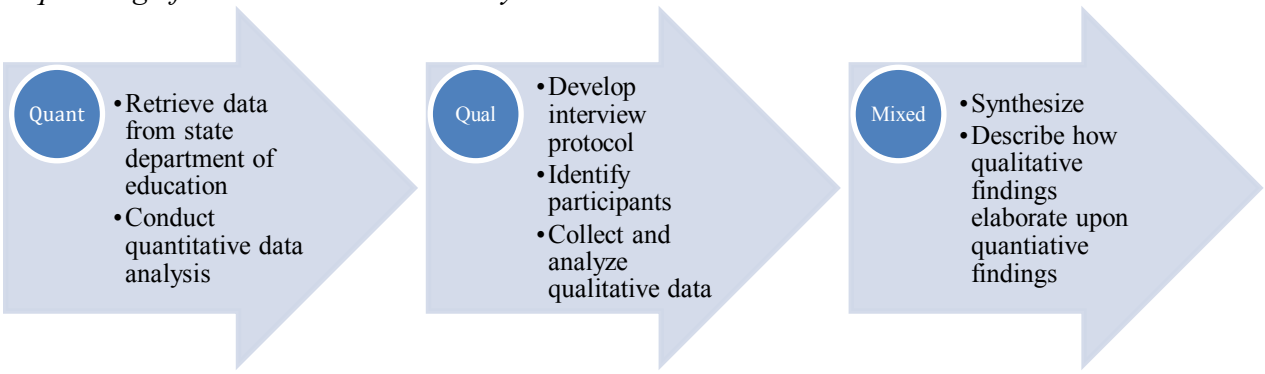
al., 2014). A statistical model that reliably predicts composite grades of school performance may be used to encourage dialogue regarding the validity of the interpretation and use of test scores. This study contributes to the ongoing dialogue of how systems of accountability, such as A-F report card grades, support valid inferences about school quality and performance.

Methods

Research Context and Design

This study used an explanatory sequential mixed-methods approach (Creswell & Plano Clark, 2017) to (1) explore the predictive ability of non-achievement-based variables on composite grades of school district performance, and (2) examine educators' perceptions of school report card grades. This research was conducted through a quantitatively driven design. The study examined a robust amount of quantitative data supplemented with a limited amount of qualitative data. The explanatory sequential design was selected with a complementarity purpose to initially explore patterns in quantitative data, and then examine educators' responses to the quantitative findings of the study. Figure 2 illustrates the sequencing of data collection and analysis. A goal of an explanatory sequential design is that the qualitative findings help to reify or elaborate quantitative findings (Creswell & Plano Clark, 2017). That is, the qualitative findings can explain the quantitative findings and offer scholars better insight into a phenomenon under investigation (Onwuegbuzie & Collins, 2007). The explanatory sequential design aligns with a complementarity purpose for conducting mixed methods research. A complementarity purpose of mixed methods research uses one strand of data to "illustrate, elaborate, or clarify the results from another strand" (McCrudden et al., 2021, p. 2). In this study, we adhere to this explanatory sequential approach by first describing the quantitative results, followed by qualitative findings, and offer an integrated view of the two outcomes.

Figure 2
Sequencing of data collection and analysis



Ohio contains a diverse set of eight distinct types of school districts broadly classified as urban, suburban, and rural, which are listed on a state-level department of education website (ODE, 2014). Quantitative data consisted of non-achievement-based data, such as social background and school district demographics, and the composite grades of school district performance that are not determined from EOG and EOC achievement tests. Qualitative data consisted of

educators' (i.e., teachers and school administrators) responses during semi-structured interviews. There were three research questions for this study.

- (RQ1) To what degree does a statistical model using non-achievement-based variables reliably predict existing composite grades of school district performance?
- (RQ2) Which of the non-achievement-based variables used in this study are most important in predicting group membership?
- (RQ3) How do educators describe the utility of school district report card grades and their perceived connections to non-achievement-based variables?

Data Collection: Quantitative

This study used the following variables as indicators of community SES: median household income, average property value per pupil, and the number and percent of students identified as economically disadvantaged. Graduation rate and "prepared for success" are the two component grades of school performance that are not determined from EOG or EOC achievement tests. The *prepared for success* component grade measures college and career readiness by collecting information about student participation in opportunities such as advanced placement courses, ACT testing, and vocational education (ODE, 2021). Data collected for this study were publicly available through ODE. The variables of interest were determined based on prior literature and existing data.

The dependent categorical variable was school districts' composite letter grade. Letter grades were collapsed into three categories. School districts earning a grade of A or B were combined into a high-performing category, and school districts earning a grade of D or F were combined into a low-performing category. This was done for two reasons. First, only 30 of the 601 school districts earned a grade of A, and only 4 of 601 school districts earned a grade of F. Second, a grade of A or B indicates a high performing district whereas a grade of D or F indicates a low performing district. Low performing districts are more likely to be subject to state-level interventions if achievement levels fail to improve. School districts earning a grade of C are described as average performing. Twenty-three continuous independent variables were considered for this study. Independent variables were non-achievement-based variables. That is, they were not directly determined from EOG or EOC achievement tests. For a complete list of variables, see Appendix A. Data from 608 public school districts were retrieved. Seven districts were removed from the study due to missing data and/or districts identified as outliers due to small enrollment numbers.

Data Collection: Qualitative

Six participants, three teachers and three school administrators, were selected to partake in semi-structured interviews. Maximal variation sampling (Creswell & Plano Clark, 2017) was used to identify participants. This sampling technique was chosen to purposefully select educators representing a variety of school districts. For this study, participants were selected so that districts predicted as low-, average-, and high-performing were represented. Table 2 displays information of participants in relation to the school district they represent. Two school districts, represented by one teacher and one administrator, were predicted to be high performing but actually earned a grade of C.

The teachers reported a range of teaching experience from 12 years to 25 years. All participants' names are pseudonyms. Teacher participants included: (a) One middle school teacher, Amy, with experience teaching language arts, social studies, and science; (b) One middle school

Table 2
Participant variation based on school district performance

	Participants	
School District Report Card Grade	Teachers	Administrators
Predicted Grade: D/F	Amy	Benjamin
Predicted Grade: C	Jay	Tanny
Predicted Grade: A/B	Trisha	Renee
Actual Grade: D/F	Amy	Benjamin
Actual Grade: C	Jay and Trisha	Tanny and Renee
Actual Grade: A/B	N/A	N/A

intervention specialist, Trisha, with experience teaching remedial mathematics and language arts; (c) One high school intervention specialist, Jay, with experience teaching remedial mathematics and language arts. The administrators reported a range of administrative experience (i.e., does not include teaching experience) ranging from 1 year to 30 years. Administrator participants were comprised of: (a) One retired school superintendent, Tanny, with additional experience as a middle school and high school building principal; (b) One elementary school assistant principal, Renee, with less than five years of administrative experience; and (c) One first-year middle school assistant principal, Benjamin, with previous experience teaching in the same district. Data were collected through semi-structured interviews that took approximately 20 minutes each. To view the interview protocol used, see Appendix B. Development of the interview protocol was informed by the findings of quantitative data analysis. The interviews were conducted face-to-face and online using ZOOM at the preference of the participant.

Data Analysis: Quantitative

A discriminant function analysis (DFA) was conducted to determine whether non-achievement-based variables could predict composite grades of school district performance. DFA was selected because the purpose of DFA is to classify participants into groups based on a set of predictors (Stevens, 2001; Tabachnick & Fidell, 2019). DFA is appropriate for this study because the variables under study align with the variable requirements needed to conduct discriminant analysis. A DFA requires a set of continuous independent variables that serve as predictors, and a categorical dependent variable that differentiates group membership (Tabachnick & Fidell, 2019). For this study, group membership was separated by composite grades of school performance (i.e., low-, average-, and high-performing districts), and the DFA sought to classify group membership using the non-achievement-based variables identified in Appendix A. Discriminant analysis generates uncorrelated discriminant functions derived from linear relationships between predictor variables to classify, or predict, group membership (Tabachnick & Fidell, 2019). The predicted classification is then compared to the actual classification of each case.

Standardized canonical coefficients and canonical loadings were analyzed to address RQ2. Canonical coefficients and canonical loadings help contextualize discriminant functions by identifying the independent variables with the strongest relationship to the discriminant function, *Standardized canonical coefficients* can be used to describe the relative contributions of independent variables to each discriminant function (Williams, 1992). The *canonical correlation* is equivalent to the correlation between the output of the discriminant function (i.e., the discriminant score) and the categories of the dependent variable (Mertler et al., 2021). Canonical coefficients with larger absolute values indicate variables that hold more weight in relation to the discriminant function. Canonical correlations closer to an absolute value of 1.0 (i.e., a perfect correlation) are more effective at classifying cases into groups (Mertler et al., 2021). Thus, both standardized canonical coefficients and canonical loadings provide insight to which independent variables are important in discriminating group membership (Tabachnick & Fidell, 2019). As previously indicated, the larger the absolute value of the canonical coefficient and/or canonical correlation, the greater the contribution of that independent variable to the discriminant function (Tabachnick & Fidell, 2019).

Data Analysis: Qualitative

Qualitative data from semi-structured interviews were transcribed and themes were drawn out through inductive coding (Creswell & Plano Clark, 2017). The team used seven steps for inductive analysis. Two researchers coded collaboratively and concurrently to maintain validity. First, the researchers became familiar with the available data for analysis. They broadly reviewed data by reading the transcribed interviews. Step two was to review interview audio recordings to clarify any ambiguity that arose during the initial review of data. The third step was making notes about participants' perceptions of school report cards based upon the available data. Step four sought to categorize these notes. Fifth, the coders discussed categories that could be revised or eliminated based upon the findings. Step six was to review the amount and quality of evidence related to each of the final categories. The final seventh step involved drawing categories into broad themes. We engaged in member-checking to promote trustworthiness among participants and confirm whether their perceptions are accurately represented (Creswell & Plano Clark, 2017). Interpretations of each interview were shared with the interviewee, and participants were given the opportunity to confirm or identify inaccuracies in the interpretation.

Results

Quantitative Results

Preliminary analysis of descriptive statistics indicated patterns in report card grades based on SES and student attendance. The average real-estate property value per pupil was $134,810 in school districts receiving a composite grade of D or F. Whereas the average property value per pupil was $201,762 in school districts receiving a composite grade of A or B. Additionally, there are notable differences in the median household income of low-, average-, and high- performing school districts. A community with a 2018 median household income of $50,000 was three standard deviations above low- and average-performing districts, but within one standard deviation of high-performing districts (i.e., districts receiving a composite grade of A or B). In other words, examination revealed all school districts with a median household income greater

than $50,000 achieved high-performing report card letter grades. Table 3, a joint display table of quantitative results, displays descriptive statistics for variables related to SES and student attendance.

Table 3
Joint Display Table of Descriptive statistics

Variable	Qualitative Group (Predicted Letter Grade)	Mean	Standard Deviation
Median Household Income	1 - D/F	$31,398	4,196
	2 - C	$35,262	4,616
	3 - A/B	$43,396	10,758
Percent of Students Economically Disadvantaged	1 - D/F	71.4%	24.46
	2 - C	44.65%	19.33
	3 - A/B	24.69%	15.02
Property Value Per Pupil	1 - D/F	$134,810	61,372
	2 - C	$173,631	84,905
	3 - A/B	$201,762	82,223
Chronic Absenteeism	1 - D/F	20.67%	8.13
	2 - C	12.35%	5.08
	3 - A/B	7.55%	3.81
Student Attendance	1 - D/F	92.7%	1.74
	2 - C	94.5%	1.10
	3 - A/B	95.5%	0.97

In consideration of RQ1, the DFA model resulted in two discriminant functions. The first function explained 87.3% of the variance, canonical $R^2 = 0.58$. The second function explained 12.7% of the variance, canonical $R^2 = 0.16$. Thus, the two functions accounted for about 58% and 16% of the total relationship between independent variables and between composite grades. In combination, the discriminant functions significantly differentiated composite grades of school performance; $\lambda = 0.354, x^2(46) = 608.85, p < .001$. Removing the first function, the second function also significantly differentiated composite grades of school performance; $\lambda = 0.835, x^2(22) = 105.79, p < .001$. Of the original grouped cases, 74.1% were correctly classified. Cases misclassified were consistently one level from the correct designation (e.g., districts predicted to be high-performing were actually average-performing). There were no instances of a high-performing district being misclassified as a low-performing district, or vice versa.

Canonical loadings and standardized canonical coefficients were analyzed in consideration of the second research question. Variables were grouped into the following categories: (a) SES indicators, (b) school attributes, (c) race/ethnicity, (d) teacher-related, (e) and enrollment information. Canonical loadings and standardized canonical coefficients indicated school

attributes and SES indicators correlated strongly with the first discriminate function. School attributes included the prepared for success component grade, attendance, and graduation rate. SES indicators correlating strongly with the first discriminate function were the percentage of students identified as economically disadvantaged and median household income (see Table 4). The second discriminant function was strongly related to students' race/ethnicity, school attributes, and median household income (see Table 5). It is interesting to note that certain variables correlate strongly with both functions but in a contrasting direction. For instance, graduation rate has a positive correlation with the first discriminant function and a negative correlation with the second discriminant function.

Table 4

Discriminant Function One: Canonical Loadings and Standardized Canonical Coefficients

		Function 1	
Variable Category	Variable Description	Canonical Loading	Standardized Canonical Coefficient
School Attributes	Prepared for success component percent	0.744	0.361
	4-year graduation rate	0.693	0.364
	Student attendance rate	0.702	0.151
	Chronic absenteeism	-0.729	0.044
SES Indicators	Percent of students economically disadvantaged	-0.744	0.276
	Median household income	0.519	-0.148

Table 5

Discriminant Function Two: Canonical Loadings and Standardized Canonical Coefficients

		Function 2	
Variable Category	Variable Description	Canonical Loading	Standardized Canonical Coefficient
School Attributes	Prepared for success component percent	0.516	0.668
	4-year graduation rate	-0.360	-0.480
	Student attendance rate	-0.069	0.668
	Chronic absenteeism	0.126	0.601
Race / Ethnicity	Percent of students identified as Black	0.548	0.608
	Percent of students identified as White	-0.576	0.223
	Number of students identified as Multiracial	0.384	0.502
SES Indicators	Median household income	0.497	0.424

Qualitative Findings

Two primary themes arose from inductive coding the qualitative data. The first theme was that report card grades promoted the comparison of school districts. Using report card grades to compare different school districts encouraged undesired competition among schools. In some cases, this led to students open enrolling in neighboring districts based on the perception that they will receive a higher quality of education. Open enrollment refers to a system where students are permitted to attend a public school district that neighbors the school district in which the student resides. Tanny explained:

> the main purpose of [the report card grades] is public relations. It's not for improving curriculum, although we would try to do that also. It's moreso for PR, parents from [low-performing schools] wanted to send their children to us, but we had parents and families trying to open enroll students to [high-performing] surrounding schools.

However, Renee noted that open enrolled, or transient, students are not always the result of differences in academic achievement across school districts. "Occasionally we have students enroll in the district because they are moving in with a different parent or a grandparent," explained Renee. That is, transient students may be experiencing unfortunate circumstances in their home-lives, such as a change in parental custody.

Participating teachers also described the use of report card grades to compare the quality of education across school districts. Trisha explained how "families that move into the area look at report card grades because they want the best place for their kids, and they pick [the affluent community] because that district earned an A." This competition based on report card grades may have additional consequences based on whether districts permit open enrollments. Jay suggested, "[The high performing districts] wouldn't allow [economically disadvantaged] individuals to attend their school because [families who are economically disadvantaged] can't afford housing there." Furthermore, participants suggested school report cards have limited capability in fostering improvements in instruction. Amy expressed the emphasis is on raising test scores, but not improving instructional practices. She stated, "There's pressure coming from administration to get scores up... [teachers] end up teaching to the test, although we're not supposed to do that, but [teachers believe they] have to do that to get scores up." That is, report card grades encourage individuals to focus on an educational product rather than the educational process. Amy's point raises concern that gains in student achievement may be the result of "teaching to the test" rather than improvements in instructional quality.

The second theme was that educators are unsurprised by the ability of non-achievement-based variables to predict report card grades. All educators accepted the evidence that non-achievement-based variables predict report card grades. "The demographics prevail," Tanny proclaimed, who also suggested that a predicted grade of C is representative of "an average income and average educated community." Tanny explained how the educational orientation of the community was generally to graduate high school, but not to pursue a college education. Educators were not surprised by the correlation between socio-economic status and report card grades. Benjamin noted:

> This is a poverty issue... schools are judged as apples to apples, and yet if we can see [SES] is a big divider in predicting student scores then maybe we change how we grade schools based on [SES] because we already know [achievement levels] will be different.

Trisha expressed sympathy for teachers of a low-performing urban school district. She stated "Year after year they have poor report card grades and get put on academic watch... I feel bad for them because I think those teachers are trying really hard, but those poor kids, their home lives, they don't get the help they need." Additionally, many teachers expressed concern about student achievement data collected as a single data point. Amy and Jay both suggested using a locally controlled assessment system to measure academic growth. Both teachers argued a locally controlled vendor assessment is a better representation of student learning compared to a single achievement test.

Integrated Findings

The utility of the A-F report card accountability system is the primary theme permeating the quantitative and qualitative findings. A primary purpose of educational accountability systems is to improve the quality of education being provided to students by raising standards for education (AERA et al., 2014; ESSA, 2015; Kane, 2013; NCLB, 2001; ODE, 2021). Perhaps a meaningful finding from the qualitative data is what school personnel did not say about school report card grades. School personnel did not describe school and district report card grades as the engine that drives education reform in their classroom or school district. Amy described how teachers in her district revert to teaching to the test in an effort to raise students' achievement scores. Amy even acknowledged a sort of internal struggle teachers are faced with, specifically, feeling the need to teach to the test when they recognize such a practice is not representative of high-quality instruction.

Some school personnel viewed the report card grades in a manner that provided context to our quantitative findings, that the report card grades indicate more about what students bring to school rather than the performance of schools themselves. For example, Tanny's interpretation of her district earning a letter grade of C on their report card reflected the community's SES and educational orientation. Classification results of the DFA correctly predicted Tanny's school district as a school expected to earn a letter grade of C on their report card. Benjamin also recognized the influence that out-of-school factors have on report card grades, and DFA classification results correctly predicted his school district to earn a D or F on their report card. Benjamin suggested that the DFA classification results of report card letter grades are evidence of a poverty issue rather than school performance or teacher quality. Taken collectively, we draw the conclusions that the educational accountability system described in this study lacks utility in (a) measuring the quality of education being provided to students, and (b) improving the quality of education being provided to students.

Discussion

Results from the current study demonstrate differences in non-achievement-based variables, such as SES indicators, can predict composite grades of school performance. These results add to the findings of previous research regarding student achievement. That is, there exists an abundance

of evidence over many years that student achievement and social background variables remain strongly correlated (e.g., Broer et al., 2019; Bankston & Caldas, 1997; May, 2006). This study extends beyond findings related to student achievement and report card grades by demonstrating the predictive ability of non-achievement-based variables on measures of school accountability. In sum, it is possible to classify school performance without the use of student achievement data. This is important because current systems of accountability require the use of student-level achievement data. Perhaps the disruption to accountability systems caused by the COVID-19 pandemic provides an opportunity to consider how school districts can be held accountable.

Educators participating in this study raised concerns about the utility of school report card grades. Teachers and administrators recognized how report card grades can be used as a tool to compare the quality of education provided by different school districts, but such a comparison may not be valid when there are significant differences in students' social backgrounds. The correlation between academic achievement and socioeconomic status is well researched and documented; however, the current study plays an important role in bringing attention back to the inequities permeating education. As stated by May (2006),

> If we, as a nation, were to overtly acknowledge that wealth, or lack thereof plays a role in the success one is able to achieve, we would also have to acknowledge that some individuals are privileged by wealth and may even be bestowed with such at birth. (p. 52)

For instance, significant discrepancies in median household income separate low-, average-, and high-performing school districts. Thus, the A-F report card system perpetuates the notion that wealth is a key factor in obtaining access to a high-quality education.

The validity of accountability systems that rely heavily on achievement test results, such as the A-F report card, merit close examination. Kane (2013) explains, "Extrapolations to different kinds of performance in various contexts rely on empirical evidence" (p. 15). This empirical study presents validity evidence, based on relations to other variables, that does not support the use of student achievement test results to measure school performance. Results from this study suggest non-achievement-based variables, such as indicators of socioeconomic status, influence measures of school accountability. Meanwhile, there is a lack of evidence from previous empirical studies to support the claim that greater student achievement is indicative of higher quality education (Wiliam, 2010). The educators participating in this study clearly described how they view the utility of report card grades - as primarily a tool to compare school districts. It is noteworthy to acknowledge what educators did not say about report card grades in comparison to the purposes of report card grades according to ODE. More specifically, educators did not describe how report card grades (1) depict school progress in raising achievement, (2) provide information about the strengths and weaknesses of school performance, or (3) accurately represent the quality of education being provided to students.

Limitations and Future Study

The current study was limited by the availability of existing data. Data were publicly available through the state department of education. Although the findings from this study raise concerns regarding the utility of school accountability systems, specifically those systems using A-F letter

grades, this study only examined data from a single state. Generalizability is limited to some degree when considering states with significantly different demographics and states with different educational accountability systems. Future quantitative studies may consider the predictive ability of variables such as school climate and student engagement. Future qualitative studies may collect data from a larger sample of educators and/or samples of parents and community members to explore how additional educational stakeholders perceive the utility of report card grades. The voice of educational stakeholders is important in explaining the utility of educational accountability systems. A limitation of the current study is the small sample of educators that were interviewed. From a policy perspective, this study may be used to encourage dialogue regarding the validity of school accountability systems, and how those systems may evolve to better represent the quality of education being provided to students.

Results from this study do not imply that using student achievement data to measure school performance and quality is inherently invalid. Such a conclusion would require a systematic review of the evidence for and against using achievement test scores to measure school quality and performance. However, measures of accountability should reflect the aspects of education that educators can control (Adams et al., 2016). Developers of, and those mandating, accountability systems are responsible for the validation of using interpretations of test results to measure school performance and quality (AERA et al., 2014). Furthermore, validation is an ongoing process (AERA et al., 2014; Cronbach, 1988). A validity argument in support of the interpretation of test scores for an intended use "encompasses evidence gathered from new studies and evidence available from earlier reported research" (AERA et al., 2014, p. 21). As such, future research will play an important role in the validation of using achievement test results to measure school performance and quality.

Future research studies could seek to examine whether improvements in students' achievement, and subsequent improvements in measures of school accountability, correlate with improvements in independent measures of instructional quality. Further research is also warranted to explore the predictive bias of standardized test results, and subsequent accountability measures of school performance. This study presents evidence of patterns of associations in composite grades of school district performance. School attributes, SES indicators, and race/ethnicity were most important in discriminating school districts' report card letter grades. The correlation between SES indicators and accountability measures, and the correlation between race/ethnicity and accountability measures raise concern about bias potentially drawn from inferences of school accountability systems. As such, SES and race/ethnicity both require further examination.

Author Notes

Timothy D. Folger is a doctoral candidate in the School of Educational Foundations, Leadership and Policy at Bowling Green State University. He spent six years as a middle school math teacher in Northwest Ohio. His research interests focus on exploring validity and validation issues in educational assessment and evaluation, such as the development and use of large-scale achievement tests for educational accountability purposes.

Audrey Conway Roberts is an Assistant Professor in the School of Educational Foundations, Leadership and Policy at Bowling Green State University. She earned her Ph.D. in Education

Sciences from the University of Kentucky in 2019. Her research interests focus on using quantitative measurement in evaluation and survey research so that data users can make informed decisions with valid results.

Jonathan D. Bostic is a Professor of Mathematics Education in the School of Teaching and Learning at Bowling Green State University. His primary area of scholarship is exploring validity issues and trends within the context of assessment in mathematics education. This includes scholarship focused on developing and evaluating classroom-based and researcher-focused instruments that impact outcomes related to mathematics topics.

Correspondence concerning this article should be addressed to Timothy Folger at tdfolge@bgsu.edu

This manuscript is based in part upon work supported by the National Science Foundation (NSF #1920619; #1920621). Any opinions, ideas, or findings expressed by the authors do not necessarily reflect the views of the National Science Foundation.

References

Adams, C. M., Forsyth P. B., Ware, J., Mwavita, M., Barnes, L. L., & Khojasteh, J. (2016). An empirical test of Oklahoma's A-F school grades. *Education Policy Analysis Archives 24*(4), 1- 29. https://doi.org/10.14507/epaa.v24.2127

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

American Psychological Association. (2021). *Socioeconomic status.* Retrieved from http://www.apa.org/topics/socioeconomicstatus/

Amrein-Beardsley, A., & Close, K. (2019). Teacher-level value-added models on trial: Empirical and pragmatic issues of concern across five court cases. *Educational Policy*, *35*(6), 866-907. https://doi.org/10.1177/0895904819843593.

Arizona Department of Education (2021). A-F Accountability 101. Retrieved from https://www.azed.gov/sites/default/files/2021/04/A-F%20Accountability%20101%20Guidebook%20FY21.pdf

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37-65. https://doi.org/10.3102/10769986029001037

Broer, M., Bai, Y., & Fonseca, F. (2019). *Socioeconomic inequality and educational outcomes: Evidence from twenty years of TIMSS.* Springer. https://doi.org/10.1007/978-3-030-11991-1

Caldas, S. J., & Bankston, C. III. (1997). Effect of school population socioeconomic status on individual academic achievement. *The Journal of Educational Research, 90*(5), 269–277. https://doi.org/10.1080/00220671.1997.10544583

Cambium Assessment (2020, September). *Annual technical report: Ohio's state tests in English language arts, mathematics, science, and social studies.* Ohio Department of Education. https://oh-ost.portal.cambiumast.com/-/media/project/client-portals/ohio-ost/pdf/2017q1/ost_annual_technical_report_spring2020.pdf

Chmielewski, A. K. (2019). The global increase in the socioeconomic achievement gap, 1964 to 2015. *American Sociological Review*, *84*(3), 517-544. https://doi.org/10.1177/0003122419847165

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.

Cronbach, L. J. (1988). *Five perspectives on validity argument*. In H. Wainer & H. Braun (Eds.), Test validity (pp. 3-17). Erlbaum.

Every Student Succeeds Act, 20 U.S.C. § 6301 (2015). https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.2307/23353796

Kane, M. T. (2020). Validity studies commentary. *Educational Assessment, 25*(1), 83-89. https://doi.org/10.1080/10627197.2019.1702465

Lake, R., & Worthen, M. (2021) *State accountability systems in the COVID era and beyond*. Center on Reinventing Public Education. https://www.crpe.org/sites/default/files/v3_accountability_report_2021.pdf

May, J. J. (2006). The role of money, race, and politics in the accountability challenge. *Journal of Urban Learning, Teaching, and Research,* 2, 46-55.

McCrudden, M. T., Marchand, G., & Schutz, P. A. (2021). Joint displays for mixed methods research in psychology. *Methods in Psychology*, *5*, https://doi.org/10.1016/j.metip.2021.100067

Mertler, C. A., Reinhart, R. V., LaVenia, K. N. (2021). *Advanced and multivariate statistical methods: Practical application and interpretation* (7th ed.). Routledge.

Murray, K., & Howe, K. R. (2017). Neglecting democracy in education policy: A-F school report card accountability systems. *Education Policy Analysis Archives, 25*(109). http://dx.doi.org/10.14507/epaa.25.3017

National Council of Teachers of Mathematics (2014). *Principles to actions: Ensuring mathematical success for all.*

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115 Stat. 1425. (2002). http://www.ed.gov/legislation/ESEA02/

Ohio Department of Education (December, 2014). *2013 school district typology overview.* https://education.ohio.gov/getattachment/Topics/Data/Frequently-Requested-Data/Typology-of-Ohio-School-Districts/One-Page-Overview-of-2013-School-District-Typology.pdf.aspx?lang=en-US

Ohio Department of Education (2021). *Report card resources*. https://education.ohio.gov/Topics/Data/Report-Card-Resources

Ohio Department of Education (2020). *Technical documentation on component grades.* https://education.ohio.gov/getattachment/Topics/Data/Report-Card-Resources/Sections/General-Report-Card-Information/Component-Grades-Technical-Documentation-1.pdf.aspx?lang=en-US

Ohio Revised Code § 3302.03 (2018 & rev. 2022). https://codes.ohio.gov/ohio-revised-code/section-3302.03/11-2-2018

Onwuegbuzie, A. J., & Collins, K. M. (2007). A typology of mixed methods sampling designs in social science research. *The Qualitative Report, 12*(2), 281-316.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, *29*(1), 103-116. https://doi.org/10.3102/10769986029001103

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, 8(3), 299-311. https://doi.org/10.1007/BF00973726

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, *12*(3), 247-256. https://doi.org/10.1023/A:1008067210518

Stevens, J. (2001). *Applied multivariate statistics for the social sciences* (4th ed.). Lawrence Erlbaum Associates.

Tabachnick, B. G., & Fidell, L.S. (2019). *Using multivariate statistics* (7th ed.). Pearson.

Utah State Board of Education. (2020). Utah Accountability Technical Manual. Retrieved from https://www.schools.utah.gov/file/ba4f83a5-0537-4f0c-b0c4-4a34c8e7c9aa

Wiliam, D. (2010). Standardized testing and school accountability. *Educational Psychologist*, *45*(2), 107-122. https://doi.org/10.1080/00461521003703060

Williams, F. (1992). *Reasoning with statistics: How to read quantitative research* (4th ed.). Harcourt Brace Jovanovich.

Appendix A

	Description of Variables Used
Dependent Variable	Composite grade of school district performance (low-, average-, high-performing)
Independent Variables	Median household income
	Number of students identified as economically disadvantaged
	Percent of students identified as economically disadvantaged
	4-Year graduation rate
	5-Year graduation rate
	Prepared for success percent score
	Student attendance rate
	Chronic absenteeism rate
	Percent of students residing within the district enrolled at the district
	Percent of students residing within the district open enrolled elsewhere
	Percent of students residing within the district attending a community school
	Average property value per pupil
	Average teacher salary
	Average teacher experience
	Number of students identified as Black
	Percent of students identified as Black
	Number of students identified as Hispanic
	Percent of students identified as Hispanic
	Number of students identified as Multiracial
	Percent of students identified as Multiracial
	Number of students identified as White
	Percent of students identified as White
	Total Number of enrolled students

Appendix B

For the past 20 years, school districts have been required to collect student achievement data that is used to make judgments about the quality of education being provided to students. One example is the report-card system that is used in [blinded]. School districts earn A-F letter grades based on the district's student achievement, graduation rate, etc. Prior research shows that student achievement is strongly related to students' home-lives (e.g., socio-economic status). My colleagues and I have conducted a study to examine whether information about students' home-lives and the students' communities can predict the report card letter grades. We didn't use any student achievement data (i.e., results from the state tests). We used data about students' home-life, race/ethnicity, median household income, and attendance. We found that traits such as median household income predicts the report card grade fairly well.

The purpose of this interview is to collect data about how parents, teachers, and administrators interpret (and use) school districts' report card grades. Your involvement in this study is voluntary and your responses will remain anonymous. Are you willing to participate in this study?

Questions:

1. **How do you perceive your school district?** (aka: What do you think about your school district?)
2. **The [State] Department of Education releases annual report card grades for each school district. What do you know about the report card grades?**
    a. *The purpose of those report card letter grades is < insert purpose from* **[State]** *here>.*
3. *Your district earned a grade of ___from the* **[State]** *Department of Education for the 2018-2019 school year.*
    a. **Think about your local school district. What does this letter grade mean to you?**
        i. *(For any school personnel, remind them to think about their work and not the school district where they live.) For further clarity, (think about the quality of education being provided at your school district)*
        ii. *[Potential Response - That students did or did not perform well on state tests]*
            1. **Why do you think that?** *(How so? Why is that?)*
        iii. **In what ways do you use report card grades issued by the [State] Department of Education?**
4. *The model combining social background and school district information predicted your district to earn a ___.*
    a. **What does this tell you about your school district?**
    b. **How does this inform your ideas about public K-12 education in [State]?**
5. **How do you feel about being able to use information about the community, like median household income, to predict K-12 school districts' report card grades?**
6. **What concerns, if any, do you have about the way schools are "graded" using a report card system?**
7. **What information would you want to know more about regarding the quality of K-12 public education provided at your school district?**
8. **Do you have anything else that you would like to share with me about report card grades for school districts that are issued by the [State] Department of Education?**