Analyzing & Reporting Achievement Gaps: Guidance for Minnesota Schools

Michael C. Rodriguez
with
Kyle Nickodem, José Palma, & Luke Stanke
Educational Psychology, University of Minnesota

&

Minnesota Assessment Group

January 2016

Table of Contents

The Context of Achievement Gaps – From Equality to Equity

Equality in education, careers, and other important opportunities has been a hallmark of American democracy, including ideas of pluralism, equal protection, and due process – as opposed to opportunities based on income, gender, ethnicity or race, language, or other demographic characteristics. We value equality in opportunities and access, as well as equality in process and outcomes. Access to high quality educational opportunities is a strong predictor of positive educational outcomes, including achievement, high school completion, and access to undergraduate education and more.

Many stakeholders have vested interests in the pipeline preparing skilled individuals needed to fill various occupations. These stakeholders are youth and families seeking opportunities and financial stability, happiness and wellbeing; educators and community-based organizations supporting youth, families, and communities; employers seeking highly prepared skilled employees; and government officials and policy makers interested in economic growth and development. The condition of our nation's education system has a profound impact on a number of interrelated areas. For instance, a recent publication from the Center for American Progress[1] estimated that closing educational achievement gaps between native-born white children and black and Hispanic children would result in an additional 5.8 percent, or nearly $2.3 trillion, increase in the US economy by 2050.

Although equality is a shared value in the US, achievement gaps exist through historical, generational, and systematic inequalities – suggesting that equality may not result in equitable outcomes. We recognize that equality is not the same as equity; *one size does not fit all*. Inequities in employment, income, housing, health, and other arenas, lead to education gaps emerging even before children enter kindergarten[2] and stubbornly persisting through the P-20 school system[3]. We face increasing racially-based education, housing, and economic segregation[4]. In most urban areas, we find centers of concentrated poverty. Rural areas also face forms of poverty, limited support services and resources, and segregation.

Often overlooked are the 300 American Indian reservations across the United States. Minnesota is home to over 7,050 American Indians, accounting for 1.3% of the state's population which is slightly higher than the national proportion of 1.2%[5]. In Minnesota, approximately 2.3% of students are American Indian or Alaskan Native[6].

The negative effects of concentrated poverty seem insurmountable as experienced by families and youth. Monitoring educational access, opportunities, and outcomes is an important policy tool, to evaluate the extent to which we reach the goals of greater equity, and to evaluate the extent to which public policy and practice support those same goals. Analysis of achievement gaps allows us to report on these goals and monitor our efforts to achieve equity.

The No Child Left Behind Act (ESEA 2001) created a federal accountability system mandating the measurement and reduction of achievement gaps for K-12 students in several important groups based on race/ethnicity, socio-economic status, English language proficiency status, and special education status. The accountability rules and guidelines increased awareness of achievement gaps for these groups and raised questions of how to properly measure and report on the gaps. The recent reauthorization (ESSA)[7] continues the role of measurement and testing.

Purpose of this Guidance Document

Given the national context and the use of achievement gap measures in federal and MN accountability systems, we created this guidance document as a means to develop some consensus regarding the estimation and reporting of achievement gaps based on a review of the literature and evaluation of approaches currently used in Minnesota.

To provide sound guidance, we clarify the extent to which ambiguity in our definitions and methods of identification of student groups interferes with consistent estimation and reporting. We provide methodological advice on analysis and estimation of gaps. Finally, we offer guidance on reporting and communicating results – in an effort to maximize meaning, relevance, and utility of such information.

These three sections are offered as guidance – and must be reviewed in the context of local schools and communities. It must be emphasized that there is not a singular best method of estimation and reporting for all situations. Therefore, the reasons for any decision made on how to estimate and report achievement gaps should be made explicit when disseminating information. Doing so will help clarify whether assessment scores are being appropriately interpreted and used. Transparency in process and decision making is critical.

This document was developed through the collaboration of the University of Minnesota, including faculty and students from the Department of Educational Psychology in the College of Education and Human Development and the Minnesota Assessment Group, which consists of Minnesota school district research, evaluation, and assessment professionals.

Core concepts around the issues of estimating and reporting achievement gaps were developed in workshops between December 2013 and January 2016. Early in this process, MAG members presented work from their districts to review current methods and approaches of estimation of achievement gaps and communication efforts. Additional sessions continued to uncover challenges in these areas.

This guidance document is a direct result of that work and reviews of earlier drafts involving MAG members from around the state.

MN School Accountability

On March 31, 2015, Minnesota's ESEA (commonly known as NCLB) flexibility waiver was extended through the 2018-2019 school year (USDOE, 2015). Under the flexibility waiver, Minnesota is exempt from the federal sanctions for not meeting Adequate Yearly Progress and the goal of all students meeting proficiency goals set by the original NCLB legislation.

To receive the exemptions, however, Minnesota was required to set new goals of 50% reduction in achievement gaps in reading and math for every student group by 2017.

Reduction in achievement gaps will also continue to be one of the four measures used in the Multiple Measurement Rating (MMR) given to schools and districts for the federal education accountability and monitoring system.

In addition to MMR scores reflecting the extent to which achievement gaps diminish, the Striving for the World's Best Workforce (WBW) state education accountability legislation was passed in 2013. This putputting forth a requirement for districts to close achievement gaps as a component of accountability at the state level.

The WBW legislation requires school districts to develop plans (or consolidate existing plans) to meet five educational goals.

The WBW goals are:

- school readiness
- grade-level literacy by third grade
- no achievement gaps among racial/ethnic/poverty groups
- high school completion
- college/career preparedness

Districts must also demonstrate community involvement in the preparation of the school plans and educational programming efforts, report annually back to the community, and submit annual summaries of progress to the Minnesota Department of Education.

The need to appropriately, meaningfully, and usefully estimate and report achievement gaps has never before been required at this level. The stakes assigned to closing achievement gaps are high and timely. Minnesota continues to face some of the largest achievement gaps in the country while demographic distributions and achievement targets are constantly shifting.

Many challenges present barriers to developing a consistent system of estimation and reporting, such as permeable district boundaries, dramatic shifts in student demographics, and ambiguous group definitions. Nonetheless, a standardized and consistent system of estimation and reporting is essential for appropriate interpretation of test scores and utility in making decisions.

A Working Definition of Achievement Gaps

Achievement gaps represent the condition ***where achievement can be predicted by non-academic factors, such as race, ethnicity, income, or neighborhood***.

Whereas demographics are generally outside the control of schools, families, and communities – achievement is not. If this view is taken, to close achievement gaps is to eliminate the predictive power of demographic factors.

However, this is not enough; we also need to maximize achievement outcomes for all. To be clear, the goal of *closing achievement gaps* is too limiting – a more rigorous, valuable, and perhaps necessary goal for the future of our communities is to **elevate the educational achievement of all to a high standard and provide access and opportunities to higher education and desired careers to all who seek them.**

Keep in mind that achievement gaps are estimated in reference to groups of students, not individuals. High achieving students do not necessarily come from one background while low achieving students come from another, but rather there is variation in the individual achievement of students within each group.

Whereas the trend for a group does not predetermine the achievement for all individuals belonging to that group, closing the achievement gap means that demographics should not even impact the *likelihood* of one's access to opportunities and level of achievement. Furthermore, there always may be high and low achieving individuals; after all, individual differences are a hallmark of humanity. But individual differences should not be predicted by group membership.

Validity of Score Interpretation and Use

Is growth in reading ability between 3rd and 4th grade an appropriate interpretation and use of scores from a science assessment taken in 5th grade? Clearly not.

Is it appropriate to interpret the difference between the mean score of White students and the mean score of Black students on a math assessment as the differences in the average math skill and ability of the two groups? That seems more reasonable, conditioned on our interpretation of "math skill and ability".

Is it appropriate to use that group difference in scores as part of an evaluation of the quality and equitability of teaching being provided to students within a school? If it is, how was it determined to be an appropriate interpretation and use?

Validity is the extent to which evidence and theory support the intended interpretations and uses of test scores and related data[8].

Validation is the gathering of evidence to support those interpretations and uses. There are two components of this validity conceptualization: an interpretation-use argument (IUA) and a validity argument.

The **IUA** is the articulation of the intended interpretations and uses (and intrinsic assumptions) of test scores.

The **validity argument** presents the evidence to support the IUA.

It is important to note that validity is not an either/or decision, but a matter of degree, such that deeper interpretations and higher stakes decisions require more rigorous evidence. As such, validity is not a characteristic of a test or measure, but is a characteristic of the inferences, assumptions, or uses of data (test scores). Every proposed interpretation or use requires validity evidence, such that no test or measure should be interpreted or used in unintended ways (i.e., in ways that do not have adequate evidence). Evidence for high stakes uses should be clear, consistent, and convincing.

Traditional interpretations of test scores as indicators of achievement are generally supported through the evidence collected during the test development process and psychometric evaluation of test score quality. That is, we assume that the test is an appropriate tool to make inferences about what students know and can do relative to state standards, particularly when we have content-related evidence and evidence of score precision (reliability and standard errors of measurement). We also note that significant differences in scores exist among groups of students, indicating important differences in what certain groups of students know and can do. The use of state test scores in this manner addresses many stakeholders at local, state, and federal levels. Such interpretations inform, among others, educational and policy-relevant decisions regarding instruction and resources. Score use also addresses local evaluation of educational programming and instructional practices.

Underlying the interpretation of achievement test scores as indicators of what students know and can do are important relevant assumptions. Perhaps most salient, we assume **a uniform and consistent level of access to and engagement in high quality instruction relevant to curriculum providing opportunities to learn the content embodied in the state standards**.

This set of assumptions includes elements of:

- High quality instruction based on teacher preparation and professional development, and school policies regarding instructional practices
- High quality curriculum with content relevance, representativeness of the rigor required in the standards, and meaningfulness to students
- Student engagement including the support and encouragement received in their schools, families, and communities
- Test quality indicated by content representation and the absence of construct-irrelevant characteristics and bias

The validity of these assumptions does not necessarily interfere with the inference regarding student knowledge and abilities, with the exception of test quality, but may interfere with the appropriateness of test score use.

Definitions and Identification of Student Groups

When estimating achievement gaps based on group performance, we introduce another set of assumptions regarding the definition and classification of group membership. These assumptions must be examined to appropriately and meaningfully interpret and use test scores as intended.

Race and Ethnicity

Race and ethnicity are complex human characteristics. Race has traditionally been defined in narrow terms given phenotypic and continental origin. Asian, African, European, and Indigenous Indian (across the Americas) origins broadly classify individuals in ways that suggest greater variation between groups than might exist within groups.

Similarly, because of the complexity of heritage from Latin American nations (not all Spanish speaking, such as Brazil), the federal government introduced the *Hispanic* label, which focuses on Spanish origin, whereas others use *Latino* perhaps for more linguistic relevance and inclusivity. The Latino identification, perhaps more than any other, has been known to be very heterogeneous in terms of racial background including people from Caucasian, Asian, African, and Indigenous origins. Therefore, when measuring the White-Latino achievement gap for accountability purposes, we are assuming that every student under the Latino label shares certain characteristics and experiences. In some schools and districts this may be a valid assumption whereas in others it may not.

Relevant to this discussion of achievement gaps, race implies a history of exclusion and segregation, from the history of slavery to legalized education, housing, and employment segregation prior to the Civil Rights Act (1964), and more recent limited access to worker rights and wage protections for some types of workers. A recent theoretical framework for understanding race-based cognitive disparities proposed the following connections[9]:

- Occupational segregation leads to income disparities.
- Educational segregation leads to disparities in maternal education and verbal abilities.
- Housing segregation leads to educational and occupational disparities.

The researchers proposing this model found empirical evidence consistent with these propositions and the set of mediators they identified explained the disparities in White-Black cognitive test scores. The mediators eliminating test score disparities included[10]:

Maternal Advantage

- Income, maternal education
- Maternal verbal ability

Parenting Factors

- Maternal sensitivity
- Acceptance
- Physical environment
- Learning materials

Race not only played a role in determining Maternal Advantage, but also resulted in culturally specific parenting styles. Finally, there was a direct effect of Maternal Verbal Ability on child cognitive ability, a verbal socialization effect.

Heterogeneity of groups

Racial/ethnic groups are nor homogeneous. Individuals, families, and communities are becoming increasingly multi-racial. In the 2013 Minnesota Student Survey, of over 162,000 students in grades 5, 8, 9, and 11 approximately 7.5% self-identified with multiple racial/ethnic categories and 10.4% identified with an ethnicity only (Latino, Somali, Hmong) and no racial group. Furthermore, Table 1 shows that every racial/ethnic group contains members that also consider themselves multi-racial. For example, whereas 1489 11th grade students identified as American Indian, only 18% of them identified as American Indian only.

Table 1
Racial Identification from the 2013 Minnesota Student Survey (percentages within row)

Race	Grade	American Indian	Asian	Black	Native Hawaiian	White	*N*	*% of Grade Total*
American Indian, Alaskan Native	5	**36.8**	3.0	11.0	2.8	47.1	2765	6.9
	8	**22.7**	3.3	18.7	3.4	62.2	2728	6.4
	9	**19.9**	2.9	17.8	2.8	66.0	2509	5.9
	11	**18.3**	2.6	21.2	2.2	68.2	1489	4.0
Asian	5	3.0	**77.8**	2.6	1.2	17.0	2759	6.9
	8	3.1	**72.3**	5.5	2.5	21.3	2850	6.7
	9	2.4	**75.3**	4.4	3.0	18.7	2993	7.1
	11	1.4	**80.9**	2.7	2.1	15.1	2653	7.2
Black, African, African American	5	7.7	1.8	**73.3**	1.4	16.6	3946	9.9
	8	12.7	3.9	**59.6**	2.4	29.5	4013	9.4
	9	11.9	3.5	**58.9**	2.8	30.5	3763	8.9
	11	11.2	2.5	**64.0**	2.1	27.3	2814	7.6
Native Hawaiian, Pacific Islander	5	18.7	7.9	13.6	**27.5**	40.7	418	1.0
	8	20.3	15.7	21.1	**17.2**	43.1	459	1.1
	9	14.2	18.0	20.8	**17.4**	44.9	499	1.2
	11	10.3	17.6	18.8	**22.9**	43.9	319	0.9
White	5	4.3	1.6	2.2	0.6	**88.5**	30073	75.5
	8	4.9	1.7	3.4	0.6	**87.2**	34871	81.4
	9	4.8	1.6	3.3	0.6	**87.8**	34778	82.1
	11	3.3	1.3	2.5	0.5	**90.6**	30829	83.4

Note: Values outlined in the diagonal are percent of students for a given grade reporting one race only. The sum of the % of Grade Total across the five races for a given grade will be greater than 100% because students can select multiple races.

Students were asked separately about Latino, Somali, and Hmong ethnic heritage. Table 2 includes the racial categories identified by students given their ethnic identification. For example, 2980 fifth-grade students identified as having some Latino heritage; and among those students, 42.5% reported no racial identification (whereas 11.2% also reported to be American Indian, etc.). In contrast, about 80% of Somali students also identified as Black, African, or African American, and over 90% of Hmong students also identified as Asian. But we also observe that at least some students in every ethnic group identified as part of other racial groups. There were also a very small number of students who reported as belonging to multiple ethnic groups (not reported here).

Table 2
Ethnic and Racial Identification from the 2013 Minnesota Student Survey (row percentages)

Ethnicity	Grade	American Indian	Asian	Black	Native Hawaiian	White	No Race Identified	*N*	*% of Grade Total*
Hispanic, Latina/o	5	11.2	2.2	7.8	3.4	**40.9**	**42.5**	2980	7.5
	8	11.1	2.7	10.1	3.9	**45.3**	**39.3**	3468	8.1
	9	10.9	2.6	8.9	4.3	**43.6**	**42.3**	3085	7.3
	11	8.6	2.4	7.6	2.9	**44.1**	**44.6**	2285	6.2
Somali	5	3.2	1.7	**77.4**	0.8	10.3	**9.2**	716	1.8
	8	2.3	1.7	**79.5**	1.5	9.3	**10.3**	526	1.2
	9	2.5	5.5	**81.2**	2.8	10.8	**5.7**	436	1.0
	11	2.6	3.8	**79.2**	2.3	9.0	**7.5**	346	0.9
Hmong	5	1.9	**91.1**	1.8	0.7	6.8	**2.5**	1015	2.5
	8	0.6	**93.7**	1.8	0.8	5.2	**2.1**	907	2.1
	9	1.4	**92.0**	2.2	1.6	4.1	**2.8**	1179	2.8
	11	0.4	**94.0**	1.2	1.0	2.2	**3.1**	1152	3.1

Usefulness of Racial and Ethnic Groups

Minnesota experiences strong immigration trends from areas including Latin America, Southeast Asia, and Africa. Because of the size and unique characteristics of these communities, they are important and relevant groups within the classifications typically used. For a growing number of districts, the categories used by MDE (consistent with federal guidelines) are becoming less useful or meaningful. Although classified under the racial/ethnic group of Asian/Pacific Islander, score trends for Hmong students tend to be different than students from Chinese or Korean backgrounds. Yet, these differences are not accounted for through a broad Asian class. Similarly, we have seen a substantial increase in the number of Somali students, classified as Black/African American/African by MDE. Score trends in the Somali community may vary differently than other students who also identify as Black. The degree to which this impacts interpretation and use of scores for estimating and reporting achievement gaps depends on the size and uniqueness of the communities within a given district.

Generational Status

The trend toward multiracial families and communities is tied to another important factor: generational status. In many cases, particularly regarding immigrant families and communities, generational status is perhaps more important and relevant regarding educational planning and programming. Three students who identify from the same racial/ethnic group might still have very different educational experiences because one student is a recent immigrant, one was born locally but to immigrant parents, and the family of the third student has lived in the area for multiple generations. Whether generational status will play a role in understanding and estimating achievement gaps will depend on the demographic characteristics of a given district.

⊙ **Take Away Point**	The five racial/ethnic groups used in the Minnesota and federal accountability and monitoring systems may not be representative of the demographics actually present in a given school or district. When **estimating** achievement gaps, districts should be sensitive to the possible heterogeneity within racial/ethnic groups, impact of generational status, and account for distinct communities that exist within the district. When **reporting** on achievement gaps, how groups are defined and the assumed characteristics accompanying that definition must be made explicit, as well as any distinctions not accounted for by the classification system used. To be useful to schools and communities, we must provide descriptions to capture the full variability and complexity of student characteristics.

Socio-Economic Status

We have long been dissatisfied with the use of weak indicators for Socio-Economic Status (SES) such as eligibility or enrollment in free and reduced lunch (FRL) programs. Such enrollment is a function of income, but arbitrarily dichotomizes a continuous range of income. SES, income, and eligibility for FRL are also temporal and may change over time. This is particularly impactful when looking at gap changes over time. Students who were originally classified as FRL might not be at a subsequent time point. Furthermore, the change in classification might not be due to an actual change in income, but rather family or student choice to apply for FRL.

In some communities and some families the act of asking for FRL is unpalatable or stigmatizing and so some do not apply for FRL even though they qualify. The often poor information regarding income and eligibility for FRL means the use of FRL as a proxy for SES can be biased and the inferences based on the relation between FRL and test scores may be inappropriate.

Missing SES data are also common regardless of the measure used. Families may desire to keep income information private, especially if it is at the upper or lower end of the SES spectrum. This leads to biased estimates related to income.

Another complication in the interpretation of SES as a classification category is its close alignment with race/ethnicity and location of residence. These variables present non-ignorable confounds. We recognize that Minnesota faces one of the largest income gaps in the nation, particularly between White and Black residents which creates a confound that interferes with the interpretation of differences in SES groups and differences in racial/ethnic groups. When racial, housing, employment, and economic segregation coincide, it is impossible to separate one from the other.

⦿ **Take Away Point**	Income differs on a continuous measure, yet students are often categorized into two groups – low-income and high-income. Doing so ignores the variability of incomes within the groups which can lead to inappropriate interpretations of score differences. Therefore, it is important that estimates of achievement gaps be sensitive to potential misclassification, systematic missing data related to SES, and accounting for variables possibly confounding with SES.

English Language Learner Status

Researchers have long argued that using ELL status as a category for the purpose of accountability monitoring is an important tool for improving equity in access, opportunities, and outcomes for a group of students historically ignored. However, there are a number of difficulties hindering accurate measurement of achievement of ELL students[11]. First, students classified as ELL are heterogeneous in a number of dimensions including level of English proficiency, parents' level of English proficiency, native language, proficiency in native language, country of origin, and previous formal education. Consequently, inferences drawn from test scores might not be appropriately generalizable to all students classified as ELL.

Another difficulty is that there is variation in the consistency and accuracy with which students are identified as ELL. In some areas ELL status is a temporary category for students who, with increasing English proficiency, are then reclassified as non-ELL. In other areas, ELL status is classified in terms of current ELL, previous ELL, and non-ELL (the student never received ELL services). Additionally, because students who reach English language proficiency are able to access the general curriculum and are no longer classified as ELL, the ELL group tends to remain less-proficient over time, including the most recent immigrants who keep this classification a moving target.

A third issue is, as many researchers have claimed, that achievement tests are unduly influenced by language proficiency (while not necessarily a measure of language proficiency), thus underestimating the true academic knowledge of ELL students. Although the onus of building evidence for a test measuring knowledge equally across all student groups is on test developers, it is an issue that those estimating and reporting on achievement gaps must consider.

As found in the case of low-SES students, ELL status presents a confound in many cases, particularly among communities where immigration is more common, with large numbers of Southeast Asian, African, and Latino students. In some (rural) schools, all ELL students are also Latino students resulting in complete overlap, counting these students twice in accountability metrics. Educational researchers provided an early critique of the NCLB regulations regarding this point, noting that schools with higher levels of diversity face multiple instances of jeopardy. For instance, for some schools in Minnesota, if they receive a low achievement gap reduction score in the MMR calculations for failing to close the achievement gap for ELL students, they are likely to also receive a low score for failure to close the gaps for other student groups, such as Latino students or students from families with low-incomes.

Researchers have also provided important insights and tools to more effectively meet the instructional and learning needs of ELL students. A recent research-to-practice brief provides a summary of some of these findings and recommendations[12].

<table>
<tr>
<td>⊙

Take

Away

Point</td>
<td>There are a number of issues that make estimating achievement, and therefore gaps in achievement, for ELL students difficult:

The heterogeneity of students classified as ELL
The inconsistency in the identification of ELL students
The possibility of inappropriate language-based influence on test performance
ELL status often confounded with racial/ethnic group and SES

When estimating and reporting on achievement gaps regarding ELL students it is important to:

Describe the demographic composition of the ELL group.
Describe how ELL students are identified and classified
This will help clarify if inferences from test scores can be appropriately generalized across the entire group and whether ELL status is likely to be confounded with another student group of interest.
Monitor performance of ELL students throughout their educational career, while receiving ELL services and after.
This avoids the problem of the moving target, where students obtaining proficiency move out of this classification and new non-English speaking students take their place.

</td>
</tr>
</table>

Special Education Status

Educators and advocates have long argued for the inclusion of students receiving special education services in accountability systems even though federal regulations required it before NCLB. Despite federal regulations, students in special education programs were often ignored and, in some cases, such programs simply provided childcare with little if any curriculum-based instruction. Consequently, estimating achievement and achievement gaps for students receiving special education services is complicated through variable access to the general curriculum. For most, integration into general education classrooms is the primary goal and services are provided to support access to the curriculum – to provide the least restrictive educational environment.

In some cases, specific learning disabilities interfere with learning in ways that limit cognitive functioning, cognitive capacity, short- and long-term memory, attention, and other characteristics that facilitate learning. Although accommodations may be provided routinely, they may not provide full access to the assessment to the same extent that they provide access to the classroom curriculum.

For students with the most significant impairments, alternate achievement standards may in fact be imbedded within Individualized Education Program (IEP) goals. These might include daily living skills at a functional level and acknowledge that some students will never live independently without significant assistance. In such cases, there may be very little access to the general curriculum. Given a wide range of levels of access and integration in the general education program, the interpretation of test scores for students with IEPs as a group becomes less appropriate and meaningful.

Research also continues to show that minority, low-SES, and ELL students are disproportionately represented in special education. While MDE has specific criteria for placement in special education, districts and schools must ensure the criteria are applied appropriately to all students. Otherwise, achievement gap estimates will not only be confounded by demographic variables, but access to the full curriculum taught by a subject matter expert may be unnecessarily limited for these students.

Although less frequent than in the case of ELL status, special education status can change over time. As students are evaluated given the goals set forth in their IEP, more or less intense services may be warranted. In the case of meeting one's IEP goals, an IEP may be closed, ending special education services.

⊙ **Take Away Point**	Special education students provide a clear example, of how achievement gaps are a byproduct of unequal access and opportunity to learn a curriculum from a qualified subject matter expert. The special education group also illustrates the wide variability of achievement within the group that must be taken into account when making inferences about student knowledge and a schools' ability to improve student knowledge. School districts should find meaningful groupings of students receiving special education services that are sensitive to the nature of the cognitive impairment and level of participation in the general curriculum.

Methodological Considerations for Estimation and Analysis

There are a number of different methods to calculate achievement gap estimates and each method has certain benefits and drawbacks regarding statistical rigor and ease of reporting. The decision to use a particular method will depend in part on the type of inferences intended to be drawn from the information. The method chosen, and the assumptions associated with the method, will be part of the validity evidence collected in order to argue for the appropriateness of the intended interpretations and uses of the estimates.

In this section attention is given to factors that can bias estimates and interpretations regardless of the estimation method chosen. There are a number of other factors that could potentially impact the analysis, but these are the more commonly encountered issues in achievement gap estimation and reporting.

Proportion above a Cut Score – Percent Proficient

One common approach to explaining achievement gaps on state tests (and tests like NAEP) is to use the reporting performance level categories, proficient or not proficient, based on a cut score and then present the percentage of students proficient in each group. On the Minnesota Comprehensive Assessments (MCAs), cut scores classify students into 4 categories based on their scale score:

- Does not Meet Standards
- Partially Meets Standards
- Meets Standards
- Exceeds Standards.

Benefits

Classifying student performance into categories simplifies the data in a manner understood by a broad audience making reporting assessment results relatively easy. Also, the school accountability system requires the monitoring and reporting of students meeting standards. The process of calculating the proportion of students who scored above or between various cut scores is also a fairly simple process that requires little knowledge of statistics.

Drawbacks

First, by categorizing scores, more sensitive information about student performance and variability is lost[13], resulting in lower statistical power, and increased Type 1 errors[14]. Second, dichotomizing data leads to a decrease in the reliability of measures[15]. Third, categorizing data can play havoc with measures of effect size, typically done in estimating achievement gaps (as shown below), therefore decreasing the usefulness of the simplification process[16].

Additional negative consequences of proportion above a cut score (PAC) methodology stem from the dependency on the location of the cut score on the scale. A cut score close to the mode of a group's distribution will tend to have higher trend magnitudes than if the cut score were closer to the tails of the score distribution[17]. This occurs simply because there is a higher likelihood of students crossing the cut score given the higher proportion of the students located in the middle of the distribution. As PAC is the method used by the majority of states to calculate AYP as mandated by NCLB, this has led schools and districts, intentionally or unintentionally, to focus resources on students close to the cut score and neglect lower performing students[18].

Another way PAC results are dependent on cut score location is that gap estimates will increase if the cut score is located closer to the mode of the higher achieving group or decrease if the cut score is closer to the mode of the lower performing group. Note that in this instance, the distributions of the two groups don't change, only the location of the proficiency cut score. This means any changes in standard setting is likely to result in different PAC achievement gap estimates independent of actual student performance.

Example and Implications

Categorizing or dichotomizing data into performance levels might make some audiences more receptive to the results of complex data, the negative consequences, including inconsistent conclusions about the data, clearly outweigh most benefits.

Consider the following example of two schools. For a particular student group, School A has an average reading scale score of 48, and 35% are proficient. For school B, the average scale score is 41, but has a proficiency rate of 50%.

School	Average Score	% Proficient
A	48	35
B	41	50

School B has a higher proficiency rate because of a large spread in scales scores for students, whereas scores have very little spread in School A. If we examined proficiency rates, we might conclude that School B is a better performing school, but we know that students in School A have a higher average scale score.

This example demonstrates how the use of PAC (percent proficient) hides achievement patterns by failing to account for variability in scores which can result in grouping and classification error. Grouping error describes when students who are different in achievement are grouped in the same category, or those with similar achievement levels are grouped in different categories.

For example, one student scores a 340 on the MCA reading test and is grouped in the Partially Meets category which ranges from 340 through 349. Another student who obtains a score of 349 is also grouped in the Partially Meets category. The standard error of measurement (SEM) for these scores is 5 points, indicating that a student with a 339 is more similar to the student scoring 340 (in two different categories) than is that student with the one scoring 349 (in the same category). Given the SEM of the scale score, students near the cut scores of 340 and 350 might be improperly classified. In this case the student with a score of 349 might belong to the "Meets Standards" category, whereas the student who scored 340 might actually belong to the "Does not Meet" category.

⦿ **Take Away Point**	Results of state achievement tests may yield valid interpretations of student knowledge and skills; however, when scores are categorized and aggregated they ignore score variability leading to numerous potential errors, and inferences become dependent on cut score location rather than the actual distribution of scores. Although reporting the proportion of students who have reached a proficiency level might be easily understood by the public, the methodology is lacking the validity evidence required to support inferences about achievement of student groups, gaps in achievement between groups, and judgments of school effectiveness[19].

Aggregated Mean Scores

The aggregated mean provides a description of the location of a score distribution with a single number that can be used to make comparisons between groups.

Benefits

The aggregated mean, typically called the mean or the average, is a commonly known and easily understood statistic even to those with little statistical knowledge. This makes it a prime candidate for reporting on group-level achievement and gaps in achievement between groups. It is more descriptive than percent proficient, since it indicates the location on the scale where the distribution of scores is located, not just the percent above a specific score point, but the location of the typical student.

Drawbacks

Using aggregated means as the sole indicator of group performance, ignores variability in score distributions that can severely impact inferences and conclusions drawn from the data. For groups that are based on smaller numbers, distributions can depart from normality and the mean becomes less descriptive of the central tendency of the distribution. When score distributions contain outliers, skewness, or other instances of non-normality, the mean score becomes an inaccurate and often misleading representation of group performance.

When a total mean value is used for comparison purposes, if the student groups being compared differ in size or composition, further distortions in score interpretation can occur. A phenomena called Simpson's Paradox is a profound example of how reliance on aggregated means can be detrimental.

Simpson's Paradox

Simpson's Paradox[20] occurs when the effect of an explanatory variable on an outcome variable is reversed when a third confounding variable is considered. When the confounding variable is omitted, student outcomes are improperly grouped and improper conclusions are made.

Simpson's Paradox can be illustrated using data from the 2013 NAEP assessment. Here, the outcome variable is average reading score, states are the explanatory variable, and race/ethnicity is the confounding variable. Minnesota 4th grade students had a mean score of 227, tied for 9th highest average scale score of the 50 states. Georgia, on the other hand, tied for 28th with a scale score of 222, as shown in Table 3. Despite the lower overall score, when disaggregated by racial/ethnic groups, we find that the mean score for every student group in GA is equal to or higher than the mean score for the corresponding group in MN. White students in GA performed equally well as White students in MN whereas Black, Hispanic, Asian/Pacific Islander, and Multiracial students in GA outperformed their counterparts in MN.

Table 3
An Example of Simpson's Paradox from 2013 NAEP 4th Grade Reading

	Average scale score (SE)		*% of the state population*	
Race/Ethnicity	Minnesota	Georgia	Minnesota	Georgia
White	233 (1.0)	233 (1.5)	72	44
Black	208 (3.2)	209 (1.7)	9	34
Hispanic	207 (3.1)	213 (2.0)	7	15
Asian/Pacific Islander	223 (5.6)	245 (3.5)	7	4
American Indian	--	--	1	--
Two or more races	215 (4.7)	223 (3.8)	3	3
Overall	227 (1.2)	222 (1.1)		

Simpson's Paradox arises because the racial/ethnic distributions of students in MN and GA are vastly different and confounded with performance. White students, on average, score substantially higher than students of color. As a result, the state mean score becomes correlated with the state's proportion of White students. Since the proportion of White students in MN is much higher than the proportion in GA, MN's mean score is weighted more heavily toward the mean score of the White students, who perform at a higher level than the Black and Hispanic students in GA comprising a substantial portion of its students.

This highlights how using aggregated mean scores to make inferences can possibly lead to drawing erroneous conclusions because mean scores hide potentially important information. Within MN, there exists districts and schools with vastly different compositions of students based on a variety of demographic variables (e.g. ELL status, SES, SPED status). If any of these demographic variables are correlated with performance, then failing to account for them when making comparisons based on aggregated mean scores can produce misleading or reversed conclusions.

⦿ **Take Away Point**	The aggregated mean is a widely known and understood description of a distribution of scores; however, it does not provide information on the variability or distribution of scores. When relying solely on means for comparing groups, this often leads to, at best, inappropriate conclusions and, at worst, inferences that are opposite from what is truly occurring. Mean scores can still be an effective way to report scores, but steps must first be taken to ensure that the mean is an appropriate and accurate representation of the relevant group performance.

Standardized Mean-Difference Effect Sizes

Effect size estimates of group mean differences account for variation in groups' score distributions and provide a standardized measure of the magnitude of differences. Cohen's *d* is a common measure of mean-difference effect size and is calculated as

$$d = \frac{(\bar{X}_1 - \bar{X}_2)}{SD}$$

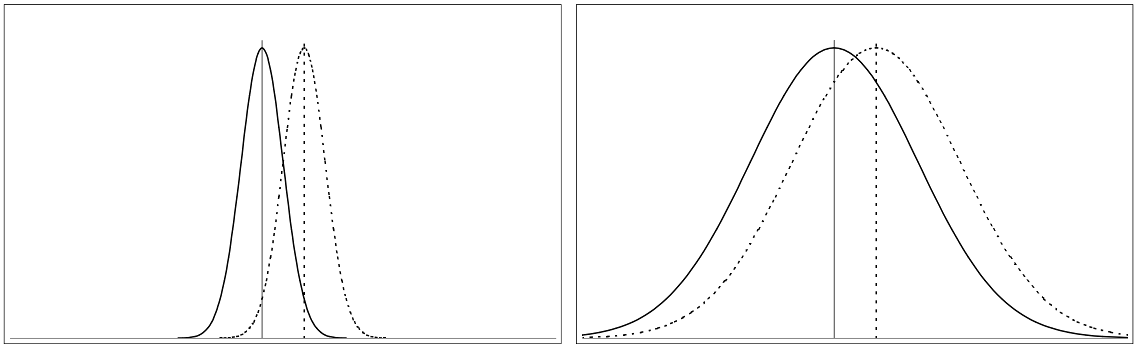
where $\bar{X}_1$and $\bar{X}_2$ are the mean scores for group 1 and group 2 respectively. The difference between these two means is then divided by a common or pooled standard deviation (*SD*). The resulting *d* statistic is measured in standard deviation units with larger values representing greater differences between the two groups. Cohen's *d* could, theoretically, take any range of values, but in education we typically interpret a *d*-value of 0.2 as a small difference, 0.5 as a medium difference, and 0.8 or higher as a large difference. An effect size of 0.8, for example, would be interpreted as a 0.8 standard deviation difference in mean scores between the groups being compared (the difference in means is as large as 0.8 of a standard deviation of scores).

Benefits

A broad audience is unlikely to be as familiar with achievement gaps reported as standardized mean difference effect sizes as compared to reporting aggregated means. Nonetheless, the interpretation that a larger effect size indicates a larger difference in scores should still be easily understood. More importantly, the use of effect sizes provides greater validity evidence for the inferences drawn from test scores than can be obtained through merely looking at aggregated mean score differences because it accounts for the variability in scores[21].

In Figure 1 below, the difference in group means in [A] is the same as the difference in [B]. Yet, in [A] the narrow distribution of scores for both groups results in less overlap in scores, and therefore, makes the difference in mean score significant. Conversely, the score distributions in [B] show considerable overlap meaning that the observed difference in scores is less significant. Effect size estimates are able to communicate the impact of score distribution whereas aggregated mean comparisons cannot.

More importantly, standardized mean differences allow for meaningful comparisons across measures or groups. It is appropriate to compare magnitudes of *d* statistics across grades (grades 3-8), test subject areas (reading versus math), or racial groups (White v. Black and White v. Latino) since the metric for each comparison is in terms of SDs.



[A] [B]

Figure 1. Influence of distributions on the significance of mean score differences.

Drawbacks

Although there are many positives to using effect sizes to estimate and report achievement gaps, it is not a flawless methodology. The pooled SD used in calculating Cohen's *d* assumes that the two groups being compared have equal population variance. If not, changes in standard deviations across multiple comparisons (e.g., across years or between various groups) can bias the estimate in favor of the group with the larger SD[22]. Therefore, it is better to use a common reference SD. For instance, when comparing racial/ethnic achievement gaps on the 3rd grade reading MCA, the SD used in all group mean-difference effect sizes should be the observed SD based on all 3rd graders, not just the SD of the two groups being compared.

Another potential issue with the use of effects sizes is that they are transformation dependent. Effect size estimations will be different if calculated with the raw mean scores as compared to scaled scores or percent proficient statistics. Although this is not inherently problematic, it is a characteristic of effect sizes that needs to be accounted for to ensure that achievement gap estimates are being calculated consistently throughout the entire analysis.

⦿ **Take Away Point**	Standardized mean-difference effect size estimates are an improvement over simple group-mean comparisons because they account for variability in score distributions and measure the magnitude of the differences in common standard deviation units. That being said, districts must still be sensitive to unequal population variances and any transformations used on the data that could impact the effect size estimates. Additionally, greater explanation of how to properly read and interpret effect sizes is likely required when reporting on achievement gaps. This is, however, a reasonable trade-off for reporting estimates that can be more appropriately used for making inferences and drawing conclusions.

Comparing Metrics and Measures

The metric chosen has implications for the meaning and interpretation of gaps. In addition, communication of differences using displays of percent proficient or mean differences could be facilitated by using a common metric like standardized mean differences, the Cohen's *d* effect size. Use of a common metric also allows for comparison of different measures, such as examining achievement gaps with NAEP versus MCAs, both in terms of percent proficient and average scale scores.

Mean Scale Score Differences on MCAs

First, consider achievement gaps on the 2013 MCA Reading and Math tests for grades 4 and 8, using Cohen's *d* standardized mean differences as defined earlier, with White student performance as the reference group. An example interpretation is Black students, on average, scored 0.81 standard deviations below White students on the 4th grade Reading MCA, which is considered a large difference. These are reported in Table 4.

Table 4

MCA 2013 Racial Achievement Gaps based on the Standardized Mean Difference d

	Reading 4	Reading 8	Math 4	Math 8
Black	-0.81	-0.78	-0.86	-0.90
Hispanic	-0.74	-0.69	-0.75	-0.74
Asian	-0.33	-0.37	-0.20	-0.10
American Indian	-0.73	-0.74	-0.74	-0.89

Proportion Proficient Differences on MCAs

An equivalent calculation can be made from percent proficient data using the probit *d* effect size for dichotomous data. Assuming an underlying normal distribution, $\boldsymbol{d}_{\text{probit}}$ will be an unbiased estimator of the population standardized mean difference in proficiency. This estimator behaved well under controlled simulation in comparison to six other estimators[23]. The standardized proportion-difference effect size ($\boldsymbol{d}_{\text{probit}}$) was computed using the Practical Meta-Analysis Effect Size Calculator[24].

The interpretation here is the proportion of proficient Black students is 0.86 standard deviations less than the proportion of proficient White students, also considered a large effect. Notice that in each case of group and test, the effect sizes are larger when using the standardized proportion difference (d_{probit}), in many cases 0.10 (one-tenth of a SD) or more. These are reported in Table 5.

Table 5

MCA 2013 Racial Achievement Gaps based on the Standardized Proportion Difference d_{probit}

	Reading 4	Reading 8	Math 4	Math 8
Black	-0.86	-0.83	-0.96	-0.97
Hispanic	-0.80	-0.78	-0.86	-0.82
Asian	-0.38	-0.43	-0.31	-0.16
American Indian	-0.86	-0.78	-0.83	-0.97

Standardized Differences in NAEP Results

Using 2013 NAEP results, a slightly different pattern is observed with respect to these same metrics. In 3 of 12 cases, the *d* standardized mean difference is larger where most differences are within 0.05, except Asian Math 4 and Hispanic Math 8. Additionally, the proportion proficient rates on the NAEP tests are substantially lower than those on the MCAs (results for American Indian students are not reported because of small sample sizes). These results are reported in Tables 6 and 7.

Table 6

NAEP 2013 Racial Achievement Gaps based on the Standardized Mean Difference d

	Reading 4	Reading 8	Math 4	Math 8
Black	-0.69	-0.84	-0.90	-1.11
Hispanic	-0.70	-0.75	-0.85	-0.76
Asian	-0.27	-0.30	-0.30	-0.27

Table 7

NAEP 2013 Racial Achievement Gaps based on the Standardized Proportion Difference d_{probit}

	Reading 4	Reading 8	Math 4	Math 8
Black	-0.73	-.89	-.91	-1.14
Hispanic	-0.66	-.74	-.85	-.94
Asian	-0.08	-.31	-.39	-.28

Finally, there is the comparison of the magnitudes of achievement gaps between MCA and NAEP results given the two metrics (*d* and d_{probit}) through which to estimate gaps. Figure 2 illustrates this for 4th Grade Reading results. Notice the gaps are larger for MCAs than they are for NAEP.

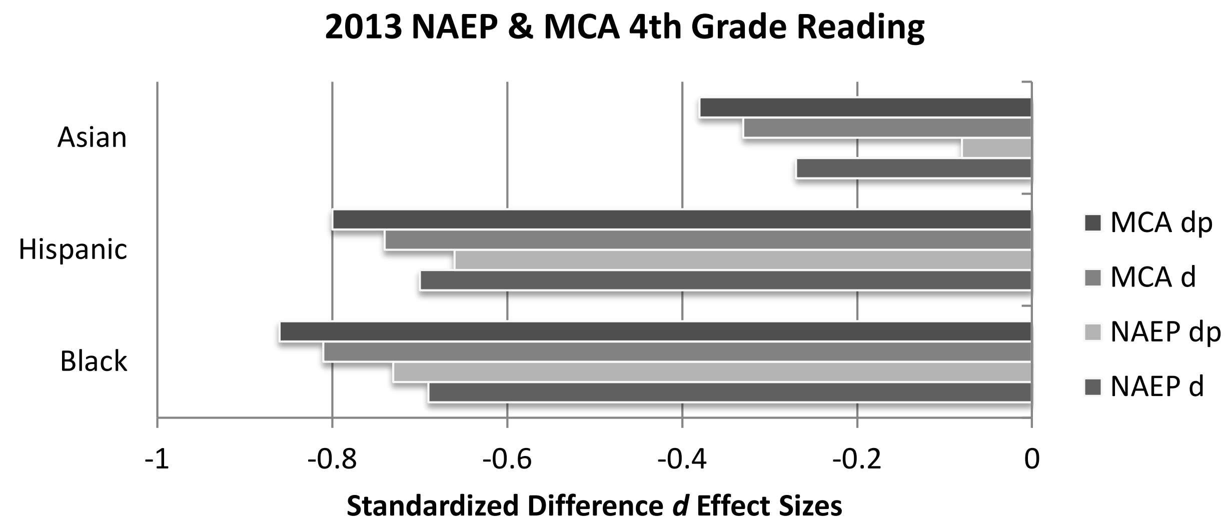


Figure 2. Comparison of NAEP and MCA 4th grade reading achievement gaps by standardized mean (*d*) and proficiency rate (*dp*) difference effect sizes.

Alternatively, consider Grade 8 results in Science, Math, and Reading, as three different subject areas. Figure 3 illustrates the achievement gaps in 2013 MCA results based on standardized mean differences (*d*) from White students as the reference group; Figure 4 contains the same results for NAEP.

In this case, some estimates of achievement gaps are larger for NAEP results than MCAs. Also, for most groups and on both measures, the gaps observed in Science are the largest (except for American Indian students where they appear to be about the same).

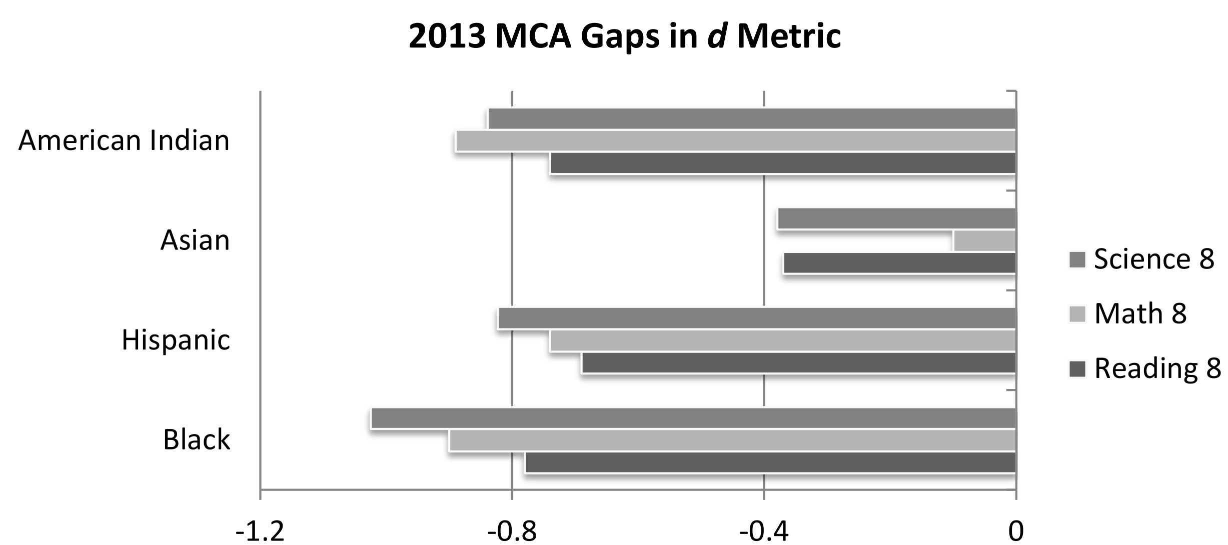


Figure 3. Comparison of MCA 8th grade achievement gaps in Reading, Math, and Science standardized mean difference (*d*) effect sizes.

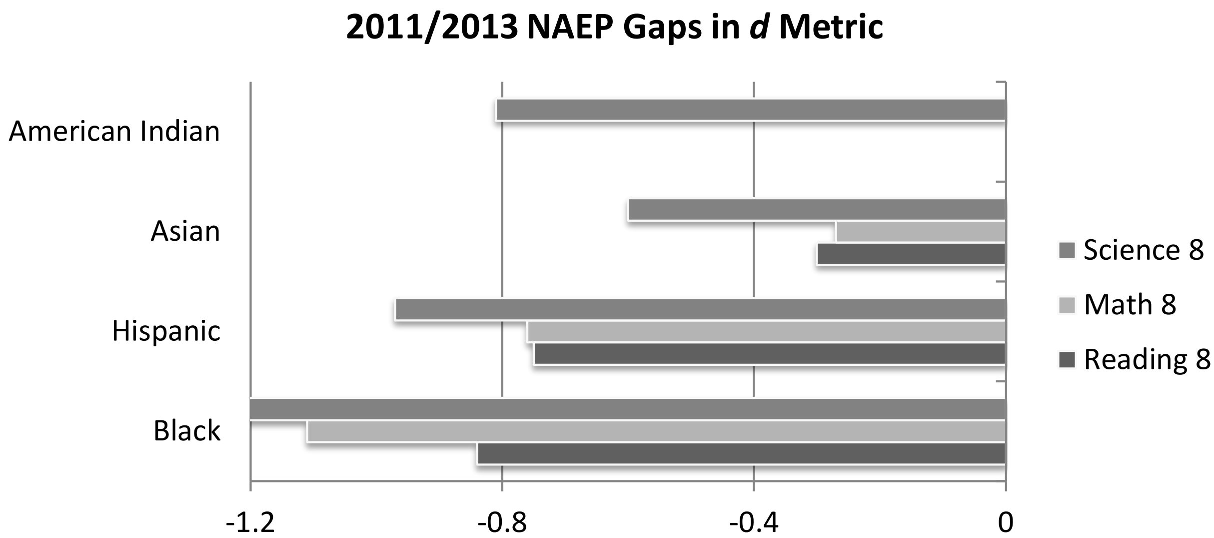


Figure 4. Comparison of NAEP 8th grade achievement gaps in Reading, Math, and Science standardized mean difference (*d*) effect sizes. Science scores are from 2011.

Finally, trends are also influenced by choice of metric and measure. Figure 5 illustrates the variation in change in achievement gaps over time given two different measures (MCA and NAEP) and two metrics (*d* for mean differences and *dp* for proportion proficient differences). Very different change profiles are observed for the achievement gaps for Minnesota Black students based on different measures (MCA and NAEP) and metrics (mean score and proportion proficient).

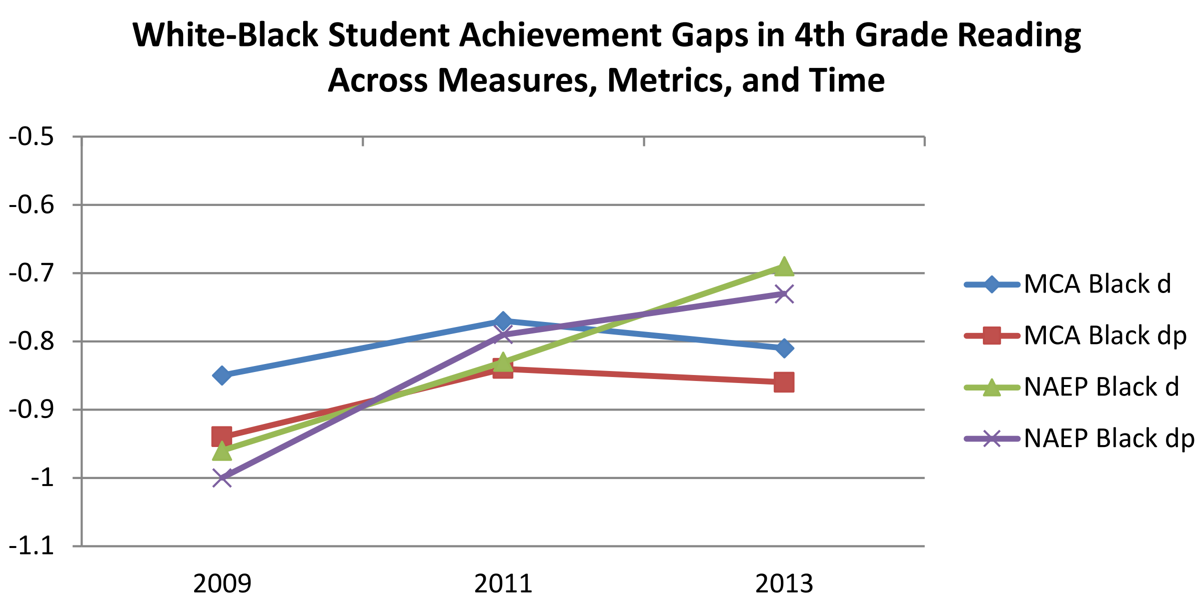


Figure 5. Comparison of NAEP and MCA 4th grade reading achievement gaps over time by standardized mean (*d*) and proficiency rate (*dp*) effect sizes.

All of the values used to create the displays in this section can be found in a companion document created for MAG, reporting 2013 NAEP and MCA achievement gaps by race/ethnicity for MN students[25].

⦿ **Take Away Point**	Different measures and different metrics may provide dissimilar estimates and interpretations of achievement gaps between groups and over time. Conversely, if multiple metrics and measures produce similar interpretations, reporting the estimates from multiple sources provides additional support for conclusions concerning achievement gaps. Regardless of the conclusion, it is important to describe to the audience the reason for choosing a certain metric or measure and how it facilitates estimation and comparison of gaps.

Cross-Sectional v. Longitudinal Designs

The design of an achievement gap study has an impact on the interpretations and inferences that can be made about gaps. With cross-sectional designs, data are collected at a single time point. In education this often means for a given school year, students across all grade levels are assessed.

Cross-sectional designs allow inferences to be made about achievement gaps between groups at the time of measurement. They do not, however, allow for inferences to be made about achievement gap trends – how gaps might change over time. There are three approaches to longitudinal study design: trend, cohort, and panel designs.

Trend Design

Trend design is simply the use of multiple cross-sectional measurements, such as the typical reporting of 3rd grade proficiency rates over time. Although trend designs are often used to make inferences about changes in achievement gaps, these types of inferences are inappropriate because they rely on the assumption that 3rd grade students are from the same population each year – or that students entering 3rd grade begin with the same level of prior achievement or skill sets and are as equally prepared for learning as prior 3rd graders. In reality, not only does the proportion of students within each demographic group vary from year to year, but it is impossible to account for the variation in students' educational experiences that could explain changes in the observed achievement gaps. Unfortunately, it is common for these types of inferences to be made from cross-sectional trend designs, often leading to erroneous conclusions.

To illustrate the inappropriateness of using multiple cross-sectional measurements to identify trends, especially compared to the conclusions produced by the more methodologically sound cohort and panel designs, we use an example from the MCA-II Reading scores for grades 3-7 from 2008 to 2012 for students in Minneapolis Public Schools (MPS) who identified as Black and White[26]. Table 8 contains mean score differences between Black and White students by grade and by year. The value in the Change column represents the change in the Black-White achievement gap for each grade between 2008 and 2012. Based on the results, if looking at cross-sectional measurement trends, we would conclude that in five years MPS has made only small or negligible reductions in the Black-White achievement gap with the exception of 6th grade. However, such a conclusion relies on the often violated assumptions described in the previous paragraph. When compared to the conclusion produced by the more sound methods presented below, the danger of using trend design becomes clear.

Table 8
Black-White Gaps in MCA-II Reading Score Means by Grade and Year

Grade	2008	2009	2010	2011	2012	Change
3	-25	-26	-29	-26	-24	**1**
4	-22	-23	-22	-22	-22	**0**
5	-21	-21	-20	-18	-19	**2**
6	-22	-20	-21	-19	-17	**5**
7	-20	-24	-22	-20	-19	**1**

Note: Change is the difference between the achievement gap in 2012 and 2008 for each grade.

Cohort Design

There is a significant amount of information that trend analysis does not take into account because it attempts to make comparisons between distinctly different groups of students (different populations) measured at different times. A slightly better, although not optimal, approach if only cross-sectional data are available is to look at how achievement gaps change for cohorts of students as they move across grades.

Cohort analysis involves following a group over time, but the group membership changes. If the entire population of students at a given grade is the group, change in membership could be for any number of reasons, including student mobility or dropouts. If only using a sample of students each year, the group membership will change as a result of re-sampling each year. Using the same data from Table 8, compare the gap differences diagonally in Table 9 as opposed to horizontally. This results in a different picture of the gap trends.

Students who were in 3rd grade in 2008, the cohort for whom we have the most data, started with a Black-White gap of -25, but the gap was reduced by two points in 4th grade, three points in 5th grade, and one point in 6th grade while holding steady in 7th grade; between 3rd and 7th grade, the Black-White achievement gap had been reduced by 6 points.

Table 9
Black-White Gap in MCA-II Reading Score Means by Cohort and Year

Grade	2008	2009	2010	2011	2012	Change
3	**-25**	-26	-29	-26	-24	
4	-22	**-23**	-22	-22	-22	
5	-21	-21	**-20**	-18	-19	4
6	-22	-20	-21	**-19**	-17	10
7	-20	-24	-22	-20	**-19**	9
Change			-2	-1	2	**6**

Note: Change is the difference between the achievement gap in a cohort of students' most recent year of data as compared to their most distant year (following cohorts over time diagonally).

While this approach is superior to looking at a single grade-trend over time, it still has a number of limitations that impact the types of inferences that can be made. Although students tend to move from one grade to the next together, it is by no means exactly the same group. As discussed later in this document, students who dropout tend to be low-performing and minority, thus inflating the mean score for minority groups and artificially closing the achievement gap in later grades. Furthermore, student mobility can impact estimates in any number of ways that cannot be predicted, in terms of students entering and leaving a school or district. Lastly, even if the estimates are accurate, this approach still cannot provide any explanation for why the gaps changed. Without knowing precisely which students were included in the estimates, it is impossible to account for differences in student educational experiences, and there are many.

Panel Design

Longitudinal panel designs, on the other hand, follow a fixed group of students over time. This means assessing the same group of students each year for a number of years. In doing so, **panel designs allow direct inferences to be made about the change in achievement gaps over time.** Particularly, by tracking the same students and gaining additional information about each student's educational experience, panel designs allow for variation in scores to be attributed to certain shared or unique experiences.

Minneapolis Pubic Schools[27] REA staff conducted a longitudinal study with a student panel design who began 3rd grade in 2008, which aligns with the data used in the previous examples. By tracking the student panel each year from 2008 through 2012 as the students completed 7th grade, the study found that the Black-White achievement gap in reading scores was reduced from -25 to -17, an eight point reduction, illustrated in Figure 7.

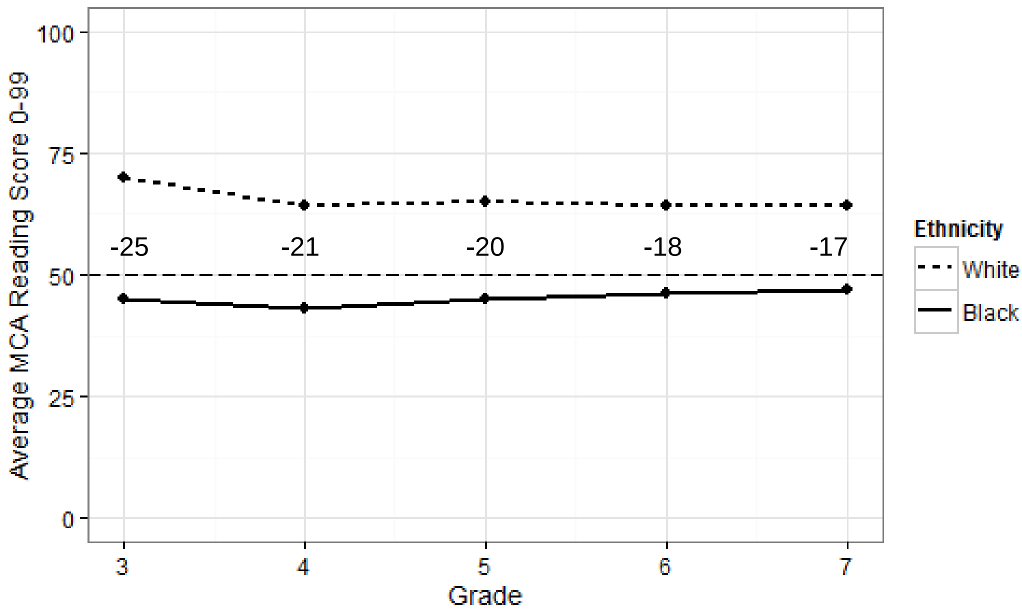


Figure 6. Average MCA-II reading scores for Black and White students from a 2008-2012 longitudinal panel cohort.

By measuring the same students every year, longitudinal panel designs are able to provide more accurate estimates of change in achievement gaps over time because they are not unduly influenced by changes in group membership. In this case, the longitudinal design revealed that achievement gaps in MPS showed greater reduction than what was inferred from the trend or cohort designs. Another advantage of panel designs is that reduction or expansion of gaps can be attributable to particular grades.

In Figure 7, although the mean reading score for Black students shows a gradual upward trend over time, there is a bit of a dip in 4th grade. Thus, MPS may want to investigate what occurred in 4th grade that accounts for this dip. In addition, notice that part of the reduction in the gap is due to a decreasing average score for White students. It is only through longitudinal panel designs that these types of inferences can be made.

The information presented in the example above has at least two potential limitations. First, attrition continues to be a challenge. Students may leave a school or district, reducing the number of students in a panel over time. There are two ways to deal with this. One is to begin with the specific students available in the most recent year (7th grade) and pull their data for previous years – so there is no change in student membership of the panel over time. Another option is to evaluate the difference in scores for student retained versus students who left to test whether there is a self-selection bias or difference in students who remain over time.

Finally, these types of displays suggest that the point estimates are precise and without error, and that the scores are comparable over time (common scale). When panel analyses are conducted and the final group of students might be a sample of the beginning group of students (due to attrition over time), some indication of sampling error might be informative, by including error bars on the point estimates. For most analyses, schools and districts have population data – so no sampling error bars are necessary. But in some cases, as in the earlier discussion of generalizability over other characteristics, error bars might facilitate inferences beyond a particular group of students at a particular point in time. Moreover, it might be useful to include error bars on point estimates that convey information about variability – to illustrate spread around a point. In these cases, error bars could represent a single standard deviation, to remind readers or audiences that there is variability, and in some groups, more variability than others.

Finally, the MCA scores across grades are from different tests on different scales. Although the reporting metric (G50, where G=grade) is common, each grade-based test is scaled independently. Because of this, different levels of variability in scores might indicate lesser or greater differences across grades. One way to account for variation in score scales is to use the standardized mean difference *d* statistic, which adjusts for differences in score distributions across measures (grades) and time (years). This would provide a much stronger common metric for evaluating changes in achievement gaps for student panels as they progress through grades.

State proficiency rates in Minnesota, for example, have shifted dramatically over the last 10 years as tests have changed to match new and more rigorous standards. This, in turn, can also skew changes in achievement gaps across these administrations due to shifting cut scores and range of measurement on the state tests, as well as opportunities to learn the new content.

⦿ **Take Away Point**	Cross-sectional designs are useful for describing score differences between groups at a given time; however, they are unable to provide information about trends over time that can be interpreted with certainty and clarity. In order to make inferences about changes in achievement gaps over time, a longitudinal design is required. Different longitudinal designs offer different levels of rigor in supporting strong inferences about changes in achievement gaps over time – from trend designs and cohort designs to panel designs, whereby the most rigorous design is the use of the same panel of students measured at each time point.

Other Factors to Consider

Changes to the Standards/Assessments

Minnesota set new mathematics standards in 2007 and reading standards in 2010. These were accompanied by new assessments in 2011 and 2013 respectively, to properly measure the content in the new standards. When this occurs it is inappropriate to make comparisons from the old to new assessments because they are based on different content.

State proficiency rates in Minnesota, for example, have shifted dramatically over the last 10 years as tests have changed to match new and more rigorous standards. This, in turn, can skew changes in achievement gaps across these administrations due to shifting cut scores on the state tests, as well as opportunities to learn the new content.

Student Mobility

In 2013, 13% of students statewide changed schools mid-year with the rate varying widely across districts[28]. Mobile students are defined as the students who have transferred in or out of a school during a single year. At least three factors are closely associated with student mobility: family income, population density, and home ownership[29]. Although often associated with a residential move, it is estimated that 30-40% of school changes are related to other factors such as school environment, suspension and expulsion policies, and academic rigor[30]. Regardless of the reason for mobility, there are numerous studies negatively linking student mobility with academic achievement. Students that are highly mobile are more likely to be below grade level in both math and reading[31].

Mobility also affects more than just student achievement. Mobility is also negatively associated with attendance and school climate[32] and excess staff stress[33], though positively related to dropout rate[34] and remediation-related costs[35].

Student mobility complicates the inferences that can be made from achievement gap estimates. **When a student leaves one school to attend another, it is nearly impossible to account for the effect of each school on student performance and growth over time.** In some cases, the school in which a student takes the MCA is responsible for the proficiency and growth of the mobile student, even if the student may have attended the school for less than 30 days. Consequently, depending on the intended interpretation and use of scores, additional validity evidence might be necessary before appropriate inferences can be drawn. This could include investigating the primary reasons accounting for the high mobility in the school and the length of time a student has attended the school prior to taking the MCA. Of course, the MN school accountability metrics are based on the presence of students in the testing school on October 1; which helps ameliorate *some* of the temporal concerns of mobile students.

Dropout Rate

Students who dropout are typically facing persistent academic challenges. Furthermore, dropout rates have been predominately higher among underrepresented populations (another form of achievement gaps is the graduation gap). As a result, this can lead to underestimating achievement gaps and overestimating school effectiveness for later secondary school grades by being unable to include these students in gap estimate calculations. As dropouts occur over school years, the student groups become more and more selective. In other words, low performing students self-select to dropout at a much higher rate. Before inferences about achievement gaps can be made about schools and districts with high dropout rates, group dropout patterns should be examined. If no group differences are observed in dropout rates, as unlikely as that may be, this provides validity evidence for the interpretation and use of scores in the later grades. If different patterns are observed across groups, these must be taken into account to support appropriate inferences. Particularly, one must ask:

- What would achievement gaps and gap trends look like if more students from the dropout population were retained?
- Would gaps remain the same?

Group Size

The size of a student group will influence the precision with which achievement gap estimates can be calculated and the perceived change in gap scores. As an example, in 2012 there were 10 7th grade students who identified as American Indian in District A. On the mathematics MCA, 50% of the American Indian students were proficient with a mean score of 750. In 2013, with these 10 American Indian students now in 8th grade, their mean mathematics MCA score is 850, but 90% of students are proficient. The mean score suggests consistent growth relative to state standards for students as a group between 7th and 8th grade, but if looking only at percent proficient, this looks like quite an impressive improvement.

Such an instance occurs if student scores were close to the cut score in 7th grade and they performed similarly in 8th grade, but where most were just above the cut score. However, with four students scoring above the cut, we see a 40% increase in proficiency. Comparatively, if the same situation occurred in District B with 100 American Indian students, four students would yield only a 4% increase in percent proficient. This leads to dramatically different interpretations of change in achievement between Districts.

Other estimation methods can be similarly impacted by small group sizes. If a higher performing American Indian student moves to District A and scores an 890 on the 8th grade mathematics MCA, her score alone could raise the mean group score from 850 to 854. Likewise, if calculating the standardized mean difference in comparison to White students whose group mean is 855 and a state SD of 14, the gap shrinks from -0.36 SDs ([850-855]/14) to -0.07 SDs ([854-855]/14). The interpretation of the achievement gap changes from a moderate gap to almost no gap because of the addition of one student.

With small populations, variation in a single student can dramatically alter achievement gap estimates and the resulting inferences. Schools, districts, and states often do not report estimates for groups with less than 10, 20, or sometimes even 40 students (as with NAEP), as well as to protect student identity. When building evidence for the interpretation and use argument, we must consider if a group is large enough to make appropriate conclusions about the aggregated scores.

⊙ **Take Away Point**	A number of factors can unduly influence achievement gap estimates and lead to inappropriate interpretations. In instances of high student mobility, it is difficult to determine which school is responsible for the performance of the mobile student. Similarly, high dropout rates can lead to inaccurate representation of the full student population, particularly when monitoring cohorts over time. Therefore, high mobility and dropout rates should be a signal that additional information is needed to provide relevant context. Also recognize that small group sizes can result in highly unstable gaps and trends. Ultimately, sensitivity checks need to be conducted to investigate the presence of such factors and great caution should be taken when interpreting and using achievement gap estimates when they exist.

Reporting and Communicating Results

Taking direction from NCLB requirements, the practice of many states and school districts is to report achievement gaps in the form of state assessment proficiency rates across a five or 10 year span. This is driven by the goal of all students being proficient. However, if other goals are important, such as the growth of all students, then information describing score distributions over time is necessary.

The primary objective is to match the estimation and reporting methods to the purpose and goals of communicating student achievement information.

Beyond considerations of the appropriateness of various methods of estimation, which are presented above, additional factors should be taken into account when determining how to best communicate about achievement gaps, including:

- Audience
- Purpose
- Complexity
- Creativity
- Measurement limitations
- Use of multiple measures
- School and district cultures
- Community cultures.

Audience and Purpose

Each school, school district, and community has a wide range of expertise – or lack thereof – when it comes to understanding educational measurement concepts. Some stakeholders, such as parents or the school board, will take what is reported at face value and trust the conclusions presented to them regarding achievement gaps and student performance, regardless of the measurements used. Conversely, teachers, administrators, and others who often have more at stake with what is reported may desire a more in-depth explanation of how the estimates were produced. When it comes to audiences, the most important considerations are their level of expertise and related abilities to understand what is being presented, the level of trust the audience has in the presenter and what is being presented, and their need for the information and justification or deeper explanation.

Additionally, the level to which we want these stakeholders to understand what we present – our purpose – may influence the measures and analyses we choose and how we present them. Our purposes may range from tracking gaps for school accountability reporting or to engage in a research effort to understand how gaps are composed and to identify related correlates. If we have a trusting lay audience and only desire for them to understand basic ideas about gaps (i.e. that they exist, that they have been around for a while, etc.), the data presented should be untransformed and the presentation should be as direct as possible. This could be accomplished with visual displays of average performance or percentages in each proficiency level, minimizing the use of statistical jargon and calculations or additional transformations. In cases where the audience is likely to believe whatever is presented, taking the messages literally, and this cannot be overstated, **it is our responsibility to be accurate and transparent with the information presented regardless of the estimation methods used.**

This means that even though percent proficient may be most easily understood, the use of percent proficient must be evaluated to determine whether it is an appropriate, meaningful, and useful manner to report results for a specific student group, school, or district – the role of context, when relevant, cannot be avoided to develop a simple message – particularly when that message is obscured by keeping it simple. Parsimony is an important goal to strive for in research and statistical modeling; however, if parsimony interferes with a more comprehensive understanding of the complex contexts surrounding achievement gaps, it fails in terms of accuracy and transparency.

If, on the other hand, we have an audience that is not trusting or we need them to understand more complex aspects of gaps, it is important that we take the time to provide adequate background and training to ensure that everyone in the audience can reach the level of comprehension necessary. A more in-depth explanation includes explicitly describing the assumptions inherent in the analysis, the measures used and how the estimates were determined, and the measurement limitations that impact how we can interpret and use the information. In this scenario, the audience should have a clear understanding of the path that led to the conclusions presented.

⦿ **Take Away Point**	When determining how to report on achievement gaps, the following questions should be taken into consideration: • Who is my *audience*? • *What is their area and level of expertise (education, statistics knowledge, etc….)?* • *How trusting is the audience of the presenter and the information presented?* • *What is my purpose for reporting? What do I need the audience to know?* Regardless of the answer to these questions and the methods chosen to deliver the information, steps need to be taken to ensure that the information presented is appropriate for the intended interpretations and uses.

Complexity

One reason we often choose to communicate state proficiency rates over time is that it is a simple measure, commonly understood and trusted by most stakeholders. However, this does not mean that it is the best metric for the purposes of improving practice and policy.

An important consideration when determining what metric to use and report is the level of precision required in the analysis and the ability to convey information about variability and changes in variability over time.

Precision

Proficiency is not a precise measure of student achievement – it is a course categorization of an underlying continuum. Proficiency helps communicate whether students have attained a specific level of achievement on the scale of measurement, but it does not communicate information about the absolute levels of student achievement or the variability in achievement. In other words, it does not give the full picture of what students know and can do, only that they are on one side or the other of one point on the scale. If our purpose is to communicate, with some precision, how gaps in achievement are changing over time, examining changes in proficiency rates will not capture changes that occur on either side of the proficiency cut score.

If the audience and purpose of reporting suggest that presenting more complex metrics is unsuitable, one alternative is to back up simple measures with complex ones that are not presented. For example, we might look at scale scores across different versions of the test over time and more deeply explain variation in the simple proficiency rates. We may make the decision to share the gaps in proficiency rates across administrations, then augment this information with statements or graphical displays of how groups differ on the score scale. For example, we can report increasing (or decreasing) scores within each proficiency level, rather than only report the percent in each level.

Take the opportunity to educate stakeholders on the nature of the complexity, being direct about the potential misleading nature of simple analyses and offering to provide the more complex analyses with those who desire to see it. We do not do our audience justice (students, communities, educators, school leaders, etc.) when we obscure information under the excuse of complexity.

Supporting Indicators

Although this does not increase precision in any way, an alternative to reporting on test score gaps is to tie achievement gaps to other meaningful indicators. Graduation, for example, is a desired student outcome for school districts and an important component of the MMR and WBW. Although graduation rates are subject to policy changes, such as requiring (or abandoning the requirement of) graduation assessments (e.g. GRAD), they offer a measure of how well schools and districts are meeting this goal. Other factors include dropout rate, college enrollment rate, availability of college-planning support, attendance, student discipline, or measures of social-emotional learning available through the Minnesota Student Survey[36].

Information from these indicators can be used to supplement achievement gap estimates from test scores to provide converging evidence towards a given conclusion. As with gap estimates from test scores though, the collection and calculation of values from these related variables must be appropriate for the intended purposes.

Indicators such as these speak to the success of students as they find their way through the educational pipeline from early childhood through postsecondary education. We need to attend to the full range of educational successes from cradle to career.

Variability

Monitoring variability over time is also an important metric relevant to school-based efforts to address achievement gaps. As noted earlier, individual differences will always exist, but since a major component of individual differences is a function of group membership, this component of variability is inequitable and the target of significant reform efforts (total variance is the sum of variance within and between groups).

Schools that face some of the most significant challenges in achievement gaps also see greater variability in student achievement.

Clearly, the teacher that enters the classroom where students vary more in what they know and can do will face significantly more complex challenges than the teacher where students are relatively homogenous in achievement.

Consider the magnitudes of MCA standard deviations (SD) in three contexts, the statewide SD and those in Minneapolis Public Schools and St. Paul Public Schools (Table 10). Observe that the variability in these two urban districts is larger than what we find across the state as a whole – increasing the challenge facing schools in meeting the educational needs of all students.

Table 10
2013 6th Grade Mathematics & Reading MCA Standard Deviations

	State SD	MPS SD	SPPS SD
Mathematics	13.5	16.9	15.1
Reading	17.5	21.2	18.6

This is such an important area, it is treated more in depth in the following section of this document. There, preliminary results of an ongoing study, in cooperation with MDE, are reported. Reports of this work will be available mid-2016[37].

⊙ **Take Away Point**	In addition to being attuned to the purpose and audience of communication efforts, consider the desired and appropriate level of precision. Simple measures, such as proficiency rate, lack precision. Therefore, when a more nuanced gap analysis is preferred, a more complex estimation method is necessary. To support communication of complex information, consider providing information about supporting additional important indicators at the same time. In addition, consider providing information about variability of student achievement, both within and between relevant groups of students.

Creativity

One way to help stakeholders with limited knowledge of statistics and educational measurement to understand concepts of statistical and practical significance, as well as simpler ideas like group size, is to use a combination of graphical displays, charts, and tables. The displays of standardized mean differences, illustrating achievement disparities, are complex for those who will struggle with the effect size metric.

In order to effectively communicate the results, it might require a little creativity. Even when attempting to think outside the box, a few guiding principles must be maintained: **accuracy, parsimony, and transparency.** Even the most creative visual display is rendered useless if it leads the audience to improper conclusions or is too convoluted to be interpreted easily. Some examples are offered here, but they are by no means a comprehensive list.

Simple Tables of Differences

When looking at gaps in proficiency on a state test, we might display all gap values (i.e. differences in mean scores) by grade-level and color cells with different shades of red or green to align with specific standardized effect size differences, such as Table 11 below.

Table 11
Differences in Scale Scores between Students of Color and White Students on the 4th Grade Reading MCA

	2009	2011	2013
Black	-14	-13	-13
Hispanic	-13	-12	-12
Asian	-7	-6	-5
American Indian	-11	-12	-12
State SD	16	16	15

Note: Indicates an effect size < -0.80, between -0.80 and -0.40, and > -0.40.

We could similarly do this over time for a single grade with the first year used in the comparison as the baseline. Score differences are not comparable across grades because each grade-specific test is scaled independently (with different means and variability). When creating displays, be mindful of individuals with color blindness or other related visual challenges. Instead of using color shading for magnitudes of effect sizes or group sizes, we could alter the font size. Illustrating group size differences offers an opportunity to inform the audience how small group sizes can show big changes from year to year – small groups produce less stable results.

Reporting Odds Ratios

Odds ratios are a way to convert gaps in proficiency rates into another interpretative frame. Reporting dichotomous data in terms of odds can make the differences in proficiency rates concrete and more practical.

For example, it would lead to statements such as: "White students are 1.9 times more likely to Meet Standards than Students of Color" or something similar. However, this doesn't necessarily make the results more interpretable and may confuse the message for some stakeholders. For instance, since odds are generally associated with gaming or gambling, some may begin to think that proficiency contains an element of randomness or luck. Because of this, methods to convert proficiency rates into odds are not presented here, although techniques are readily available online[38].

Alternate Graphical Displays

If, as a district, we have established **goals for gap closure**, we might include lines on graphs that show the target, giving stakeholders an idea of progress over time. Targets provide the audience with a way to easily evaluate the metrics presented to them and when they are established by stakeholders, they make it easier for the audience to trust what is presented. This may be particularly useful in the context of World's Best Workforce reporting requirements.

Another way to creatively provide a way for the audience to evaluate gaps is to provide **normative comparisons**. For example, a district might compare district-level gaps to state gaps over time. This has an added advantage of accounting for external threats to validity, such as changes in state tests, because those changes would likely show up in the state data as well. A district could also consider illustrating gaps relative to similar districts or schools (Dr. Heistad, Bloomington Public Schools, has presented cross-district comparisons of achievement to MAG). Such comparisons help provide a reference group (as long as that reference group is appropriate and relevant) and a sense of progress, or lack thereof, by eliminating other threats to validity (e.g., "we are not like everyone else").

As a final note, Internet and software-based **data visualization tools** are rapidly improving and become very powerful and effective means of representing data in a stimulating, yet meaningful way. Some of the programs are rather intuitive whereas others have a bit of a learning curve. A few that are worth experimenting with are Tableau, Weave, and the ggplot2 package in R[39].

⊙ **Take Away Point**	In order for creativity to be helpful, it must not be taken too far. A simple rule for presenting any data is that it should not present too much information at once. Charts and tables provide an efficient way to communicate information about gaps, but too much information can make it impossible for the audience to walk away remembering what they just saw or end up misrepresenting the intended interpretations.

Measurement Limitations

Investigating and accounting for measurement limitations is of the utmost importance when estimating and communicating about achievement gaps. For trusting audiences, this will discourage misuse of the data. For untrusting audiences, this will provide credibility and transparency.

We have already discussed a number of measurement limitations, but some are worth emphasizing because they are commonly overlooked. For instance, proficiency cut scores can skew results when group mean scores are near cut scores because large numbers of students can cross over the cut score from year to year, simply due to measurement error. For audiences, this might appear as if gaps are shrinking or expanding significantly from year to year when, in actuality, comparing mean scale scores might reveal that gaps remain consistent.

Likewise, when group size is small, one or two students can make a noticeable impact on group-level data leading to improper inferences.

Sources of Error

Usually, the intent in measuring student achievement is to make statements about what students know and can do, but not restricted to the specific items on the test, the specific conditions of administration on the day of the test, or the specific characteristics of the student on that particular day. The interpretation is typically much broader, generalizing across all items that could have been asked about the domain (content standards) and over a relatively broader time period and conditions.

Test scores don't naturally contain error, in and of themselves, until we attempt to make inferences that go beyond the specifics of the testing event. When generalizations are brought into the interpretations of scores, which always happens, error is introduced into the score interpretation. Consequently, evidence must be provided to demonstrate that various sources of error do not unduly impact interpretation.

If interpretation of an assessment score is meant to generalize over items in the domain or content standards:

- Requires stability of scores over different items.
- Evidence can come from estimates of internal consistency or the reliability estimates of alternative forms of the assessment.

If interpretation of an assessment score is meant to generalize over time:

- Requires stability of scores over time.
- Evidence can come from test-retest reliability over the relevant time period. Collecting test-retest reliability evidence is often challenging because it requires the passage of time in the absence of intervention or learning.

Measurement Error

There are two types of measurement error: systematic and random. **Systematic error** stems from problems with the assessment, data collection, or scoring procedure. This includes, but is certainly not limited to:

- Test wiseness
- Limited English language proficiency
- Unfairly speeded tests
- Improper administration procedures
- Incorrect scoring key

Efforts should be taken to identify and eliminate systematic measurement error. While the onus is largely on the test developer to design a high-quality assessment and appropriate scoring procedure, it is often the responsibility of schools and districts to ensure that administration and collection of the assessments is appropriate and uniform.

Unlike systematic measurement error, **random measurement error** – those conditions that randomly fluctuate over a hypothetical set of multiple testing situations such as mood or focus - is impossible to eliminate. However, instead of ignoring it, scores must be reported in a way that acknowledges errors of measurement, such as confidence intervals associated with scores.

Point estimates, like percent proficient or mean scores, are subject to both systematic and random measurement error. Whereas random error is represented in variability in scores over repeated testing (standard errors of measurement), systematic error is captured in the hypothetical true score and it will always be present over repeated testing. Since true ability is not expected to change in a short period of time, if a student hypothetically took a test multiple times, the variance of a single student's scores over repeated testing is random measurement error variance. The standard deviation of these repeated scores (for an individual) is the standard error of measurement (SEM). Theoretically, these random errors are normally distributed with a mean of zero and a standard deviation (i.e., the SEM). The SEM is a tool to interpret the score of an individual student. Over large samples of students random measurement error tends to cancel out, but systematic error will impact the overall performance level of the group.

To illustrate the complexity of interpreting point estimates given the inevitability of measurement error, consider all 4th grade students and their MCA reading scores. The mean score can be estimated easily. If the students took the assessment again, systematic error would likely impact student scores in the same manner as the first testing. As a function of random error, it is likely that most students will not receive the same score. Of the students who originally scored at the mean of the test, some will score higher and some will score lower. For students who originally scored above the mean, some will score higher the second time, but more will score lower because the unreliability of test scores commonly produces a regression to the mean effect. Similarly, for students who originally scored below the mean, some will score lower, but more will score higher (i.e., closer to the mean). Whereas random measurement error adds variability to the scores, it does not affect the mean score.

Systematic error does affect the mean score, but its effect will not change upon repeated testing. Therefore, given similar testing conditions and a parallel form of the reading MCA, the score distribution will remain the same with the same mean and standard deviation even though nearly every student will obtain a different score.

Score Interpretation

This document focuses on guidance regarding score interpretation, from the initial considerations of estimation choices to reporting. However, in the context of measurement limitations, we can provide more direct guidance to support meaningful, appropriate, and useful score interpretation. Below, we present advice from the National Academies Board on Testing and Assessment and Minnesota's own MCA Technical Manuals. This includes two pieces of advice:

A test score should never be used in isolation for decision making at any level.

Scores should not be interpreted without consideration of relevant measurement errors.

This second point is often misunderstood and underestimated. We typically interpret test scores without considering measurement error. When informed, we include some notion of measurement error. The MCA Technical Manuals provide the standard error of measurement (SEM) for every score on each test. The MCA score reports include error bands on the strand scores. These errors of measurement reflect error due to sampling items from a larger content domain (the possible pool of items we could draw from to create a test). They reflect how much scores might vary if a student was administered a different sample of items (technically equivalent to sampling error – error due to sampling items). We really don't restrict our interpretation of scores to the specific sample of items on the test, but to the domain from which the items were selected – the content standards for the grade and subject.

Similarly, we don't restrict interpretation of scores to the specific point in time the test was administered, but over a longer period of time. We generalize test performance over time, from the point of testing well into the next year (what do students know and what can they do). However, this generalization of scores over time introduces error – error due to untested change over time. This error is not captured by the SEM reported in the Technical Manuals or score reports. Notice that the more we expect of scores and the more we generalize over conditions of measurement, we introduce more measurement error in score interpretation[40].

Additional sources of error are introduced through standard setting (setting the cut score for proficiency). All methods for setting cut-scores involve human judgement. In all cases, judges do not agree on the exact location of the cut scores. This variability introduces error in the selection of the cut score to identify proficient levels of performance. And, as new test forms are introduced from year to year, forms are equated to adjust for differences in difficulty. This equating, or statistical adjustment to test scores to make them comparable introduces error (equating error). These sources of error, due to estimating cut scores and to equating, are not captured in the SEM reported in the Technical manuals or score reports.

Sources of error impacting the precision and interpretation of test scores include:

1. Sampling items from the domain or pool of possible items.
2. Generalizing what students know and can do over time.
3. Locating the cut-score at the appropriate level of performance.
4. Equating scores on tests over time to secure score comparability.

Cautions from the National Academies

In 2009, the National Academies[41] published their letter to the US Department of Education regarding the proposed regulations for the Race to the Top initiative. The National Academies consists of the nation's leading advisors on Science, Engineering, and Medicine, and also includes the Board on Testing and Assessment (BOTA) – some of the preeminent experts and thought leaders in testing and assessment. In that letter, BOTA urged the Department to be consistent with measurement theory and particularly aware of how some regulations can prevent valid test score interpretations and uses. The letter provides reminders of core measurement principles – a few are reviewed here. Many of these comments can also be found in their report: Lessons Learned about Testing[42].

In a general response regarding the reliance on testing in educational reform:

> *A test score is an estimate rather than an exact measure of what a person knows and can do. The items on any test are a sample from some larger universe of knowledge and skills, and scores for individual students are affected by the particular questions included. A student may have done better or worse on a different sample of questions. In addition, guessing, motivation, momentary distractions, and other factors also introduce uncertainty into individual scores. When scores are averaged at the classroom, school, district, or state level, some of these sources of measurement error (e.g., guessing or momentary distractions) may average out, but other sources of error become much more salient. Average scores of groups of students are affected by exclusion and accommodation policies (e.g., for students with disabilities or English learners), retest policies for absentees, the timing of the test over the course of the school year, and by performance incentives that influence test takers' effort and motivation.* (p. 3)
>
> *We encourage the Department to pursue vigorously the use of multiple indicators of what students know and can do. A single test should not be relied on as the sole indicator of program effectiveness. This caveat applies as well to other targets of measurement, such as teacher quality and effectiveness and school progress in closing achievement gaps. Development of an appropriate system of multiple indicators involves thinking about the objectives of the system and the nature of the different information that different indicators can provide. Such a system should be constructed from a careful consideration of the complementary information that is provided by different measures.* (p. 4)

In response to the use of data to improve instruction:

> *The choice of appropriate assessments for use in instructional improvement systems is critical. Because of the extensive focus on large-scale, high-stakes, summative tests, policy makers and educators sometimes mistakenly believe that such tests are appropriate to use to provide rapid feedback to guide instruction. This is not the case.* (p. 10)

Tests that mimic the structure of large-scale, high-stakes, summative tests, which lightly sample broad domains of content taught over an extended period of time, are unlikely to provide the kind of fine-grained, diagnostic information that teachers need to guide their day-to-day instructional decisions. In addition, an attempt to use such tests to guide instruction encourages a narrow focus on the skills used in a particular test—"teaching to the test"—that can severely restrict instruction. Some topics and types of performance are more difficult to assess with largescale, high-stakes, summative tests, including the kind of extended reasoning and problem solving tasks that show that a student is able to apply concepts from a domain in a meaningful way. The use of high-stakes tests already leads to concerns about narrowing the curriculum towards the knowledge and skills that are easy to assess on such tests; it is critical that the choice of assessments for use in instructional improvement systems not reinforce the same kind of narrowing. (p. 10-11)

BOTA urges the Department to clarify that assessments that simply reproduce the formats of large-scale, high-stakes, summative tests are not sufficient for instructional improvement systems. The multiple choice format in particular lends itself more easily to measuring declarative knowledge than complex "higher-order" cognitive skills. Instructional improvement systems that rely heavily on such item formats may reinforce a tendency to narrow instruction to reflect little more than tested content and formats. (p. 11)

Cautions from the 2013-2014 Technical Manual for Minnesota's Title I and Title II Assessments

Minnesota's own MCA Technical Manual[43] provides important guidance regarding test use and score interpretation (in addition to comprehensive details regarding test development, scoring, scaling, and administration). In the Purpose statement of the manual, it asserts a purpose for the MN educational assessment program:

Improved student learning is a primary goal of any educational assessment program. This manual can help educators use test results to inform instruction, leading to improved instruction and enhanced student learning. In addition, this manual can serve as a resource for educators who are explaining assessment information to students, parents, teachers, school boards and the general public. (p. 9)

Regarding reported scores and score interpretation:

As with any large-scale assessment, the Minnesota Assessments provide a point-in-time snapshot of information regarding student achievement. For that reason, scores must be used carefully and appropriately if they are to permit valid inferences to be made about student achievement. Because all tests measure a finite set of skills with a limited set of item types, placement decisions and decisions concerning student promotion or retention should be based on multiple sources of information, including, but not limited to, test scores. (p. 72)

Regarding appropriate score use:

> *The tests in the Minnesota Assessment System are designed primarily to determine school and district accountability related to the implementation of the Minnesota standards. They are summative measures of a student's performance in a subject at one point in time. They provide a snapshot of the student's overall achievement, not a detailed accounting of the student's understanding of specific content areas defined by the standards. Test scores from Minnesota assessments, when used appropriately, can provide a basis for making valid inferences about student performance. The following list outlines some of the ways the student scores can be used.*
>
> - *Reporting results to parents of individual students*
>
> *The information can help parents begin to understand their child's academic performance as related to the Minnesota standards.*
>
> - *Evaluating student scores for placement decisions*
>
> *The information can be used to suggest areas needing further evaluation of student performance. Results can also be used to focus resources and staff on a particular group of students who appear to be struggling with the Minnesota standards. Students may also exhibit strengths or deficits in strands or substrands measured on these tests. Because the strand and substrand scores are based on small numbers of items, the scores must be used in conjunction with other performance indicators to assist schools in making placement decisions, such as whether a student should take an improvement course or be placed in a gifted or talented program.* (p. 77)

Regarding score interpretation for individual students:

> *Individual student test scores must be used in conjunction with other performance indicators to assist in making placement decisions. All decisions regarding placement and educational planning for a student should incorporate as much student data as possible.* (p. 78)

Regarding cautions for score use and measurement error:

> *When interpreting test scores, it is important to remember that test scores always contain some amount of measurement error. That is to say, test scores are not infallible measures of student characteristics. Rather, some score variation would be expected if the same student tested across occasions using equivalent forms of the test. This effect is due partly to day-to-day fluctuations in a person's mood or energy level that can affect performance and partly a consequence of the specific items contained on a particular test form the student takes. … Nevertheless, measurement error must always be considered when making score interpretations.* (p. 80)

Regarding the use of objective/strand-level information:

> *Strand or substrand level information can be useful as a preliminary survey to help identify skill areas in which further diagnosis is warranted. The standard error of measurement associated with these generally brief scales makes drawing inferences from them at the individual level very suspect; more confidence in inferences is gained when analyzing group averages. When considering data at the strand or substrand level, the error of measurement increases because the number of possible items is small. In order to provide comprehensive diagnostic data for each strand or substrand, the tests would have to be prohibitively lengthened. Once an area of possible weakness has been identified, supplementary data should be gathered to understand strengths and deficits.* (p. 81)

Test scores might indicate the need for additional assessment or evidence or information. Sometimes, such additional information can come in the form of qualitative information – potentially simply in the form of teacher or parent reports. In other cases, we could utilize classroom evidence in the form of class projects, assignments, homework, or classroom tests. For nearly any decision, multiple sources of information based on multiple types of information (quantitative and qualitative) will enhance the appropriateness of the decision.

In this way, multiple measures can reduce the impact of measurement error on any single measure. Through a combination of multiple measures, a more certain, richer description of student achievement can be produced, reducing the effects of error on the ultimate interpretations and decisions.

⊙ **Take Away Point**	Because of the many sources of error in scores and score interpretation, there is a core principle in educational and psychological measurement and a clear expectation in the *Testing Standards* (AERA, APA, & NCME, 2014), that **no decision should be based on the result of a single measure for any purpose**. In addition, both the advice from the National Academies and the Minnesota MCA Technical Manual urges score users to combine test scores with other information for nearly all potential uses.

Exploring Variability in MCA Performance

A research version of the 2013-2014 MCA student score file was obtained from MDE for the purpose of exploring variability in student performance – through a data-sharing agreement (9-29-2015) with the University of Minnesota. These data were combined with school-level data from the National Center for Education Statistics and the 2013 Minnesota Student Survey.

A report will be generated and made available to MAG when completed[44]. Some of the research questions that will be investigated in this study include:

- How much variation in student achievement is within versus between schools?
- How much variation is a function of student characteristics?
  - To what extent are factors like race, SES, LEP, or gender explaining variation?
  - Do student characteristics function the same way across schools?
- How much variation is a function of school characteristics?
  - How much does school composition matter?
  - Are there malleable school factors that explain variation in achievement?

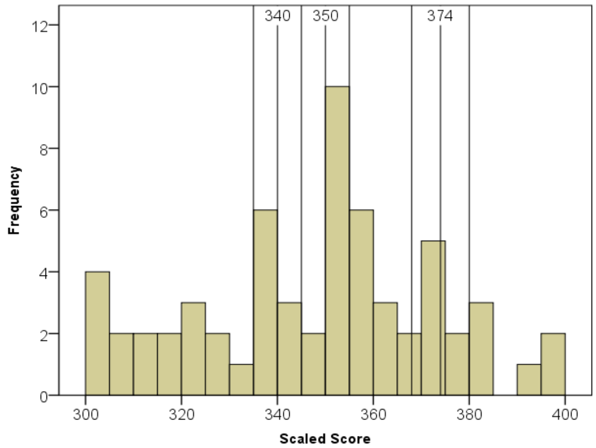
Also, with the addition of the NCES Common Core Data file with school-level information, this study will be expanded to include an analysis of the role of school composition on achievement gaps, replicating the IES study[45] of NAEP data relative to school (racial) composition. For that part of the study, the following goals will be addressed:

- Describe and explore the role of student of color density and school composition more generally
- Explore the association between school composition and achievement and achievement gaps
- Account for student characteristics and school characteristics and potential gender differences
- Explore the extent to which achievement gaps can be attributed to within-school versus between-school differences

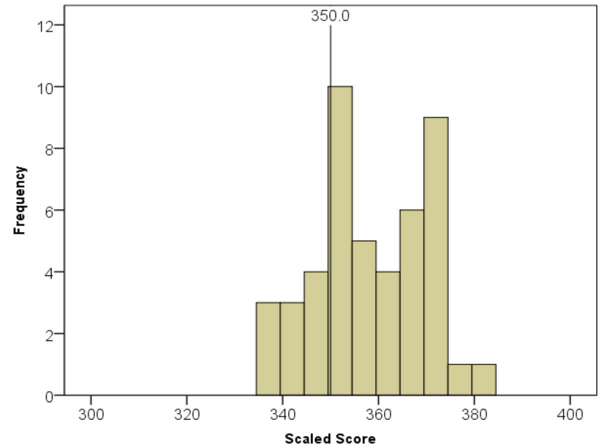
These explorations have implications for the equitable distribution or equitable use of key education resources across schools to reduce achievement gaps.

The following graphical displays are offered as examples of illustrating variability, some include information about the standard error of measurement (SEM). When there is a 1 SEM confidence interval, this is at the 68% confidence level. When there is a 2 SEM confidence interval, this is at the 95% confidence interval – recognizing that there is less precision in a score to gain greater confidence in the location of a student's score.

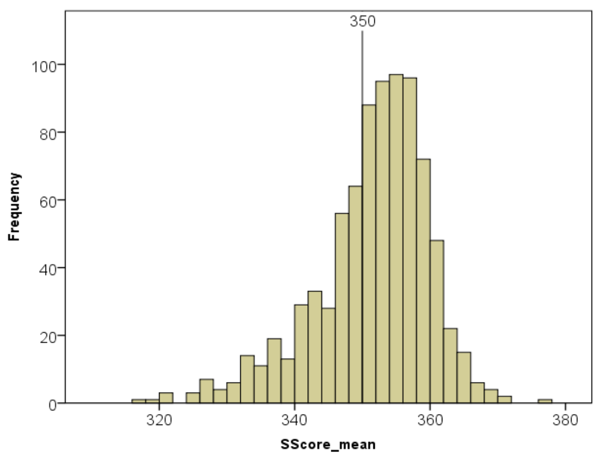
1. The first row of figures illustrates student score distributions in two hypothetical schools, which differ in terms of student variability in achievement.
2. The second row of figures illustrates the distribution of school means and school SDs.
3. The third row illustrates the school means with text-boxes locating the means of each racial/ethnic group.



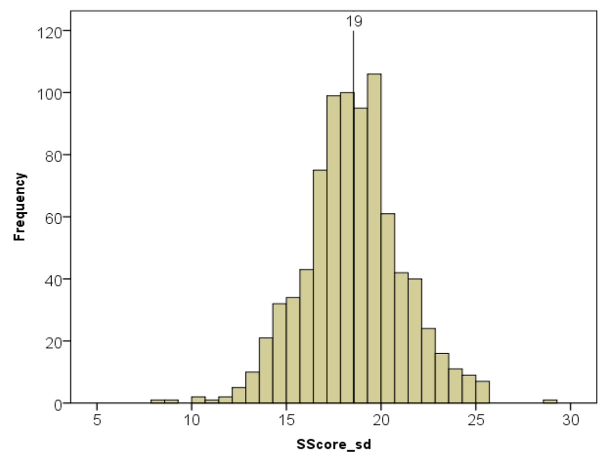
Hypothetical School A, Grade 3 Reading
n = 61, *M* = 348, *SD* = 25, ±1 SEM (68% CI)



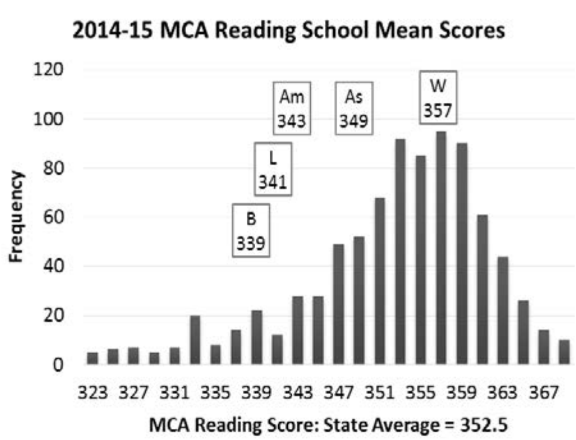
Hypothetical School C, Grade 3 Reading
n = 46, *M* = 358, *SD* = 12



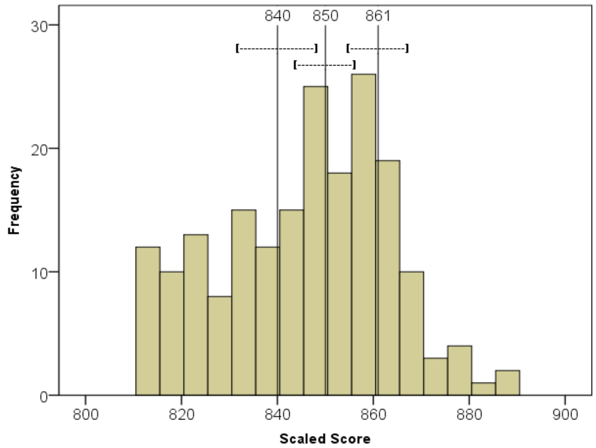
Grade 3 Reading School Means
for schools with *n* ≥ 10



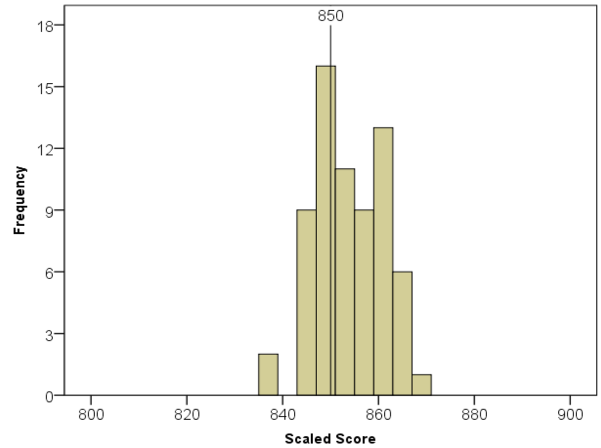
Grade 3 Reading School Standard Deviations
for schools with *n* ≥ 10



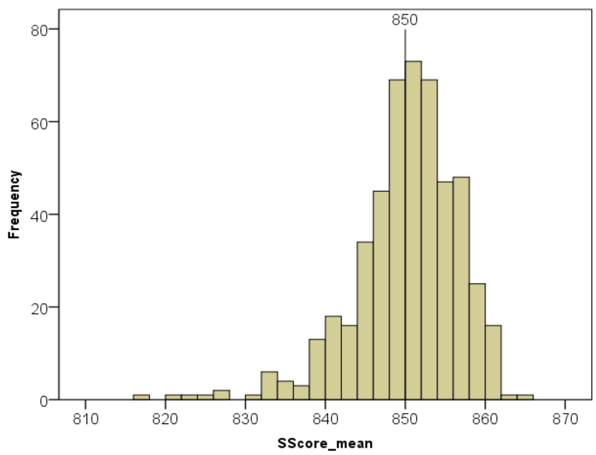
2015 MCA Grade 3 Reading
Distribution of School Means
With Race/Ethnicity Means



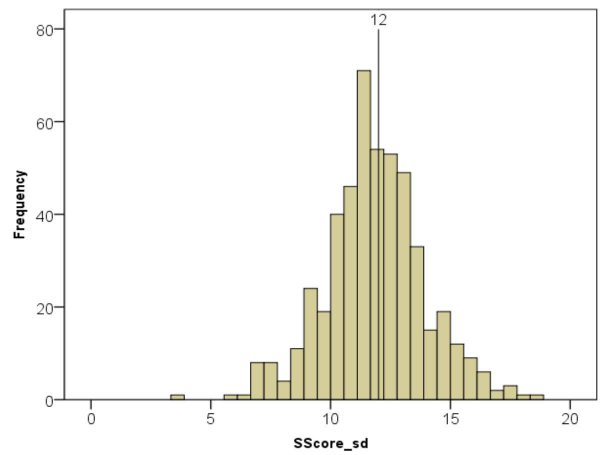
Hypothetical School A, Grade 8 Mathematics
n = 193, M = 846, SD = 18, ±2 SEM (95% CI)



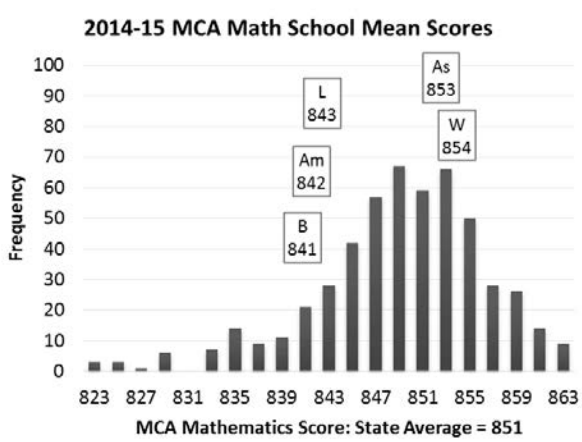
Hypothetical School B, Grade 8 Mathematics
n = 67, M = 853, SD = 7



Grade 8 Mathematics School Means
for schools with $n \geq 10$



Grade 8 Mathematics School SDs
for schools with $n \geq 10$



2015 Grade 8 Mathematics
Distribution of School Means
With Race/Ethnicity Means

The first findings from the special study of 2013-14 MCA student achievement in Minnesota allows us to report the extent to which variability in student achievement is within versus between schools. We find that a moderate amount of variation is between schools, indicating the potential for school factors to impact student achievement. For both Mathematics and Reading achievement, the proportion of variance between schools is relatively stable across grades. In addition, the proportion of variance between schools is greater for mathematics achievement than for reading achievement.

Table 12
2014 MCA Mathematics Scores – Variance Within and Between Schools

Grade	**# Students**	**# Schools**	**Mean Score**	**Variance Between**	**Variance Within**	**Proportion V(Between)**
3	59825	862	357.14	46.96	202.52	**.19**
4	60221	854	457.23	60.67	254.65	**.19**
5	58139	826	551.38	31.48	141.85	**.18**
6	56326	594	650.02	38.05	156.60	**.20**
7	57068	509	749.52	21.17	108.44	**.16**
8	55867	505	850.70	29.70	154.92	**.16**
11	52982	430	1147.75	48.96	234.05	**.17**

Table 13
2014 MCA Reading Scores – Variance Within and Between Schools

Grade	**# Students**	**# Schools**	**Mean Score**	**Variance Between**	**Variance Within**	**Proportion V(Between)**
3	59717	862	351.57	62.48	350.78	**.15**
4	60107	854	449.99	37.72	195.23	**.16**
5	57846	826	554.28	32.77	169.48	**.16**
6	56203	594	652.59	41.76	254.99	**.14**
7	57163	509	749.81	40.75	260.83	**.14**
8	56139	505	849.87	38.56	260.53	**.13**
11	55390	430	1051.66	25.95	189.30	**.12**

Note: These results include only NCES designated Type-1 (regular) schools and excludes Virtual schools.

Final Thoughts

This document provides guidance on the estimation and reporting of achievement gaps and the complex persistent challenges that schools and communities face in response. As always, educators and professionals in research, evaluation, and assessment roles have an opportunity to improve data and assessment literacy of their constituencies.

In doing so, we accept the responsibility to be informed regarding sound practice to improve data use in schools and communities and abide by professional standards for data analysis and communication. Two important sources of these standards include the *Standards for Reporting*[46] and the *Standards for Educational and Psychological Testing*[47].

Here we clarify the importance of continuing to improve data and assessment literacy in relevant constituencies, the goal of improving data use, the importance of evaluating all data-based reports for public consumption, and the need for greater understanding of validity as it relates to the evidence supporting test score interpretation and use. These components are also data/assessment literacy and use goals of Generation Next.

Data & Assessment Literacy

Data constitute the essential ingredients for evidence-based practice, quality program design, and effective policy development. Because of this, it is important that data are appropriate, meaningful, and useful. Any serious effort to address disparities in education, health, and other important arenas to promote the success of every young person, must adopt a principled approach to data collection, analysis, and reporting.

Research, evaluation, and assessment professionals can proactively enhance the understanding of data and data-based reports among all stakeholders, including policy makers, educators, community leaders, families, and youth themselves. In improving data and assessment literacy among all stakeholder communities, we can enhance the communication and understanding of the magnitudes of achievement gaps, the contexts in which they are manifested, and the monitoring of progress. In improving data and assessment literacy, we can improve data use and enhance data-driven decision making.

Data Use

To promote the effective use of data, we must understand its quality and the extent to which it is appropriate, meaningful, and useful. Research, evaluation, and assessment professionals are well positioned to evaluate the quality of data sources; ensure the appropriate, meaningful, and useful presentation of data; and monitor and promote the use and application of data and data-based reports. But to support these efforts, we must simultaneously promote and develop greater data and assessment literacy across stakeholder groups. We are driven by a set of principles and standards regarding data use and fair test use[48]. These principles inform our work and define the bases for evaluating the selection, collection, analysis, reporting, and use of data.

Evaluation to ensure Warrants & Transparency

Regarding the collection, analysis, and reporting of data, we strive to meet two fundamental requirements. These are based on the professional standards for empirical research by the American Educational Research Association and have been adopted by the Minnesota Education Equity Partnership and Generation Next to enhance equitable dissemination, interpretation, and use of research on Minnesota educational outcomes.

1. Data-based reports of findings, conditions, and change should have ***sufficient warrants***; that is, sufficient evidence should be reported to support and justify results, conclusions, and recommendations.
2. Data based reports of findings, conditions, and change should be ***transparent***; reports should clearly explain the logic of inquiry, concrete definitions of the variables or measures, the methods of data collection and data analysis, and how these result in the clearly defined outcomes as reported.

Validity & Validation

We operate on the foundational basis of validity: the degree to which evidence and theory support test score interpretation and use. Validation is the collection of relevant evidence to support score interpretation and use. In all evaluations of data reports, it is important to evaluate the utility of the reports based on the available validity evidence. We strive to maintain high expectations for ourselves in this regard and promote the use of high quality data to inform and monitor progress on school, district, and state goals (e.g., WBW). These principles should be included in efforts to improve data and assessment literacy among stakeholders. Among the many tasks and goals we face to meet these numerous responsibilities, the following are core commitments:

1. Data will be collected for clearly stated purposes and uses, for which validity evidence exists.
2. Efficiency, effectiveness, and equity principles will drive improvement efforts regarding data collection, reporting, and use.
3. Collection and reporting strategies will be periodically reviewed, on the above principles.
4. Data will be employed in a positive manner for continuous improvement, not as part of a negative campaign or for blaming.
5. Unused data will be reviewed for appropriateness, meaningfulness, and usefulness, and data collection tools will be modified if possible or eliminated.

Acknowledgements

We thank the following individuals for their contributions to this document.

NAME	District/Organization
Ashley Cozadd	Anoka-Hennepin
Johnna Rohmer-Hirt	Anoka-Hennepin
David Heistad	Bloomington
Connie Erickson	Burnsville
Donita Stepan	Byron
Jan Brunell	cmERDC
Eve Lo	Comm. Schools of Exc.
Jim Angermeyr	Consultant
Travis Voels	District 287
Evelyn Belton-Kocher	DRC
Jan Kellner	Eastern Carver Co.
Donna Roper	Eden Prairie
Susan Tennyson	Edina
Lloyd Komatsu	Forest Lake
Stephanie Strenge	Fulda
Lisa Grundstrom	Hastings
Daniel Hyson	Hiawatha Valley
Karin Swainey	ISD 197
Donna Moe	ISD 287
Paul Brashear	ISD 622
Stephanie Strenge	ISD 622
Beth Sneden	Mahtomedi
Kim O'Connor	Mahtomedi
Lynne Viker	Mahtomedi
Pam Booker	MDE
Joseph Curiel	MDE (Prev. ACET)
Amanuel Medhanie	Minneapolis
Chris Moore	Minneapolis
Eric Vanden Berk	Minneapolis
Kathryn O'Gorman	Minneapolis
Luke Stanke	Minneapolis
Tamara Weiss	Minneapolis
Matt Rega	Minnetonka
Mary Roden	Mounds View
Rick Spicuzza	Mounds View

NAME	District/Organization
Nancy DuBois	New Prague
Marie Pangerl	North Branch
Aaron Ruhland	Orono
Paul Hamilton	Orono
Don Pascoe	Osseo
Thel Kocher	Retired
Anthony Padrnos	Richfield
Cheryl Videen	Robbinsdale
Beth Sullivan	Rosemount, Apple Valley, Eagan
Michelle DeMers	
Carole Dillemuth	Roseville
Chris Bretz	Roseville
Jake Vondelinde	Roseville
Jake Von De Linde	Roseville
Dave Orlowsky	Shakopee
Melissa Miller	S. St. Paul
Chad Schmidt	S. St. Paul
Gretchen Chilkott	S. Washington Co.
Tom LaBounty	S. Washington Co.
Joe Munnich	St. Paul
John Linder	St. Paul
Stacey Gray Akyea	St. Paul
Hope Rahn	Spring Lake Park
Bob Laney	St. Anthony-New Bright.
Andrea Preppernau	St. Cloud 742
Prachee Mukherjee	St. Louis Park
Jim Potthoff	St. Michael Albertville
Dawn Madland	St. Paul Academy
Brittany Perry	Stillwater
Angie LaBounty	Stillwater
Kerry Bollman	TIES
Christine Young	Tri-City United
Peter Nelson	Wayzata
Stacey Lackner	Wayzata

Note: Some individuals have changed positions since the time of their contribution.

ENDNOTES

[1] Lynch, R.G., & Oakford, P. (2014). *The economic benefits of closing educational achievement gaps*. Washington, DC: Center for American Progress. Retrieved from https://www.americanprogress.org/

[2] Loeb, S., & Bassok, D. (2007). Early childhood and the achievement gap. In H.F. Ladd & E.B. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 517-534). New York, NY: Routledge.

[3] Davison, M.L., Seo, Y.S., Davenport Jr, E.C., Butterbaugh, D., & Davison, L.J. (2004). When do children fall behind? What can be done? *Phi Delta Kappan*, *85*(10), 752-761.

[4] Bohrnstedt, G., Kitmitto, S., Ogut, B., Sherman, D., and Chan, D. (2015). *School composition and the Black–White achievement gap* (NCES 2015-018). US Department of Education, Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch.

[5] United States Census Bureau. (2013). *Minnesota quick facts from the US Census Bureau*. Retrieved from http://quickfacts.census.gov/qfd/states/27000.html

[6] Minnesota Department of Education (2014). *Minnesota report card: Demographics*. Retrieved from http://rc.education.state.mn.us/

[7] http://www.ed.gov/essa

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

[9] Cottrell, J.M., Newman, D.A., & Roisman, G.I. (2015). Explaining the Black-White gap in cognitive test scores: Toward a theory of adverse impact. *Journal of Applied Psychology, 100*(6), 1713-1736.

[10] Ibid.

[11] Lakin, J.M., & Young, J.W. (2013). Evaluating growth for ELL students: Implications for accountability policies. *Educational Measurement: Issues and Practice: 32*(3), 11-26.

[12] http://www.collegeready.umn.edu/

[13] Altman, D.G, & Royston, P. (2006). The cost of dichotomizing continuous variables. *British Medical Journal, 332*(7549), 1080.

[14] Austin, P.C., & Brunner, L.J. (2004). Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Statistics in Medicine*, *23*(7), 1159-1178.

[15] MacCallum, R.C., Zhang, S., Preacher, K.J., Rucker, D.D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19-40.

[16] Ibid.

[17] Ho, A.D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics, 34*(2), 201-228.

[18] Choi, K., Seltzer, M., Herman, J., & Yamashiro, K. (2007). Children left behind in AYP and non-AYP schools: Using student progress and the distribution of student gains to validate AYP. *Educational Measurement: Issues and Practice*, *26*(3), 21-32.

[19] Linn, R.L. (2005). Issues in the design of accountability systems. *Uses and misuses of data for educational accountability and improvement. Chicago: National Society for the Study of Education*.

[20] Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B 13*, 238-241.

[21] Coe, R. (2002). *It's the effect size, stupid: What effect size is and why it is important*. Retrieved from http://www.leeds.ac.uk/educol/documents/00002182.htm

[22] Ho, A.D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics, 34*(2), 201-228.

[23] Sanchez-Meca, J., Marin-Martinez, F., & Chacon-Moscoso, S. (2003). Effect size indices for dichotomized outcomes in meta-analysis. *Psychological Methods, 8*(4), 448-467.

[24] Wilson, D.B. (n.d.). *Practical meta-analysis effect size calculator*. Retrieved from http://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-SMD10.php

[25] http://www.edmeasurement.net/MAG

[26] Minnesota Department of Education. (2014). *Assessment and growth files* [Data file]. Retrieved from http://w20.education.state.mn.us/MDEAnalytics/Data.jsp

[27] Medhanie, A., & Vanden Berk, E. (2013, December). *Measuring the achievement gap(s)*. Presented at the meeting of the Minnesota Assessment Group, St. Paul, MN.

[28] Minnesota Department of Education (2014). *Minnesota report card: Demographics*. Retrieved from http://rc.education.state.mn.us/

[29] Skandera, H., & Sousa, R. (2002). Mobility and the achievement gap. *Hoover Digest: Research and Opinion on Public Policy*, *3*, 1-5.

[30] Rumberger, R.W. (2002). *Student mobility and academic achievement*. ERIC Clearinghouse on Elementary and Early Childhood Education. ERIC Document No. 466 314. Retrieved from http://www.gpo.gov/fdsys/pkg/ERIC-ED466314/pdf/ERIC-ED466314.pdf

[31] United States General Accounting Office. (1994). *Elementary school children: Many change schools frequently, harming their education*. (GAO/HEHS-94-45, pp. 1–55). Washington, DC: Health, Education, and Human Services Division.

[32] Chen, G. (2008). Communities, students, schools, and school crime. *Urban Education, 43*(3), 301-318.

[33] Slater, J. (2005). Staff under strain from mobile pupils. *Times Educational Supplement, 46*(24), 16.

[34] Appalachian Regional Educational Laboratory. (2005). Student mobility and achievement. *District Administration, 41*(6), 84.

[35] Slater, J. (2005). [See full reference above]

[36] Rodriguez, M.C. (2016). *Technical report on developmental skills, supports, & challenges: 2013 Minnesota Student Survey*. University of Minnesota. Retrieved from http://www.edmeasurement.net/MAG/Rodriguez2013MSSv2.pdf

[37] http://www.edmeasurement.net/MAG

[38] A discussion of odds ratios is available in http://pareonline.net/getvn.asp?v=11&n=7

An odds ratios calculator is available at https://www.medcalc.org/calc/odds_ratio.php

[39] http://www.tableausoftware.com/public/

https://www.oicweave.org/index.php

http://ggplot2.org/

[40] Brennan, R.L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*(4), 295-317.

[41] http://www.nap.edu/catalog.php?record_id=12780

[42] http://sites.nationalacademies.org/DBASSE/Topics/DBASSE_068942.htm

[43] http://education.state.mn.us/MDE/SchSup/TestAdmin/MNTests/TechRep/

[44] http://www.edmeasurement.net/MAG

[45] Bohrnstedt, G. et al. (2015). [See full reference above]

[46] American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA Publications. *Educational Researcher, 35*(6), 33-40. Retrieved from http://www.aera.net/Portals/38/docs/12ERv35n6_Standard4Report%20.pdf

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

[48] Joint Committee on Testing Practices. (2003). *Code of fair testing practices in education*. Retrieved from http://ncme.org/resource-center

National Council on Measurement in Education. (1995). *Code of professional responsibilities in educational measurement*. Retrieved from http://ncme.org/resource-center

American Counseling Association. (2003). *Standards for qualifications of test users*. Retrieved from http://aac.ncat.edu/resources.html