What is Assessment Bias and How is Aperture Education Working to Reduce It?

Joseph L. Mahoney, Paul A. LeBuffe, Valerie B. Shapiro[1],
Jennifer L. Robitaille, Evelyn S. Johnson, and Jessica L. Adamson

APERTURE EDUCATION

ABSTRACT

Assessment bias is error that occurs consistently (i.e., not randomly), inflating or depressing the assessment scores of particular students. This document describes how Aperture Education thinks about assessment bias, why we believe it is important to address, and how we work to reduce it. We discuss how Aperture's strength-based resources, including the Devereux Student Strengths Assessment (DESSA) and the DESSA-mini, were intentionally designed to minimize four types of assessment bias from influencing DESSA scores: rater bias, construct bias, item bias, and sample bias. We also describe how assessment bias can be further reduced through specific monitoring tools, training, and implementation practices that support their proper administration, interpretation, and communication. We conclude that although it is not possible to eliminate bias in all its forms, we have a duty to work to minimize it, and the impact it can have. The field of social and emotional learning assessment is relatively new, and we plan to continue discovering new ways of identifying, understanding, and mitigating assessment bias, so that we can accurately measure and continuously improve the well-being of children, youth, and adults.

TABLE OF CONTENTS

Introduction

Aperture Education is engaged in, and committed to, studying and addressing issues related to bias, equity, inclusion, and cultural responsiveness as reflected in our resources and services. We believe social and emotional assessment should provide schools, families, and communities with data to help students and adults develop the social and emotional competencies they need to be successful in school, at home, at work, and in life. We believe these competencies develop optimally through authentic relationships that take place in safe, supportive, and culturally affirming environments. Our focus on strength-based assessment helps educators, parents, and students engage in productive and participatory planning to support the positive development of children, youth, and adults.

This document describes how Aperture Education thinks about assessment bias, why we believe it is important to address, and how we work to reduce it. We describe how the DESSA system was intentionally designed to minimize assessment bias and how the assessment bias can be further reduced through specific monitoring tools, training, and implementation practices that support its proper administration, interpretation, and communication.

What is Assessment Bias?

Assessment can be used to observe, describe, explain, and predict behavior to inform educational decision-making. Although the goal of a student behavioral assessment is to assess behavior as accurately as possible, no assessment can measure a student's behavior perfectly. Consequently, there is always some "error" influencing student scores that are byproducts of the measurement process itself. *Assessment bias is error that occurs consistently (i.e., not randomly), inflating or depressing the assessment scores of particular students.*

For example, if a specific educator views their students in an unusually favorable way, that educator may rate their students' attributes (e.g.,

conduct or achievement) more favorably than other educators rate their students – when, in fact, *there is no actual difference*. However, just because a group of students, on average, get higher or lower scores does not, by itself, mean the scale is biased. It could reflect *a real difference* (e.g., students in one classroom may, on average, have higher social and emotional competence than students in another classroom). *Our goal is to reduce the influence of assessment bias in the DESSA measures so that each rating is as accurate a reflection of student behavior as possible.*

APERTURE EDUCATION

How Do the DESSA Measures Address Different Types of Assessment Bias?

Assessment bias can have many causes that can occur in isolation or combination. Aperture Education strives to reduce or eliminate four types of bias from influencing DESSA scores: rater bias, construct bias, item bias, and sample bias. Below we define these four types of bias and describe Aperture Education's approach to evaluating and reducing them.

Rater Bias

Rater bias occurs when the person completing the rating (e.g., a parent or an educator) regularly inflates or depresses assessment scores. In other words, *the score reflects something about the person rating the student in addition to just the student's behavior.*

It is important to note that all assessments have the potential to reflect rater bias in assessment scores. Some assessments, like the DESSA, ask a rater to interpret, reflect, and judge the frequency of behaviors. Raters can be biased because of their relationship with *specific students*, their attitudes towards *specific groups* of students, or because they generally rate *all students* in a more lenient or harsh way. Rater bias can also result from a wide variety of educator or contextual characteristics (e.g., educator training; how carefully they complete the DESSA; how seriously they take social and emotional learning (SEL) in general; job satisfaction; work climate; racist, classist, or other prejudices against groups of students).

How do we Examine and Work to Reduce Rater Bias?

A combination of DESSA design characteristics, analytic approaches, and implementation practices are used to examine and reduce rater bias.

Design Characteristics. Design characteristics are properties or features of the DESSA developed intentionally to reduce rater bias. The DESSA was designed to reduce the impact of bias in four ways: the DESSA is specific, strength-based, brief, and incorporates multiple perspectives.

Specific. First, the DESSA is a *behavior rating scale* that asks raters to indicate how often the student engaged in *specific positive behaviors* over the past four weeks. This focuses the rater on what they have observed in a defined period of time, reducing the likelihood ratings will be based on an overall impression of a student that reflects non-observable behaviors or subjective judgments. Considering the behavior across four weeks also avoids focusing on a single instance of the behavior and supports a more reliable rating.

Strength-Based. Second, the DESSA is a *strength-based assessment*, meaning that the items query positive behaviors (e.g., get along with others) rather than maladaptive ones (e.g., annoy others). The ability to focus on and rate student strengths, rather than their problem behaviors, can both minimize the likelihood of items triggering bias-based beliefs and avoid deficit-oriented labeling of students. Moreover,

> **The DESSA was designed to reduce the impact of bias in four ways: the DESSA is specific, strength-based, brief, and incorporates multiple perspectives.**

asset-based framing is often more acceptable to raters and can increase their level of comfort and willingness to complete ratings in a valid and reliable way.

For example, Rosas and colleagues (2006) found greater alignment between parents and educators' ratings of a student's strength behaviors than ratings of a student's problem behaviors. Educators may be apprehensive to endorse and discuss negative items such as "fights with other children" if they must communicate or defend such ratings with parents. But, there may be less anxiety discussing a positive item (e.g., "gets along with others"), even if rated low. Moreover, parents may be more comfortable with a teacher using an educative approach, focused on building strengths and acquiring skills, rather than focusing on managing or eliminating negative behavior.

Brief. Third, the DESSA is designed to be a brief assessment. When raters are hurried or fatigued, the quality of ratings can diminish. The DESSA-mini takes about 1 minute per student, and the DESSA assessment takes about 5-8 minutes per student. The time it takes each rater to complete the DESSA is recorded, allowing rating time to be a study variable from which to examine potential bias. Ratings made hastily are of particular interest because haste may signify carelessness or the influence of stereotypes. To prevent this, raters are automatically notified if their scores were completed unusually fast and invited to review the scores before submitting.

Multiple perspectives. Fourth, we recognize that gathering multiple perspectives on student behavior, across various contexts, can be important. Each rater has a distinct and valuable point of view. Therefore, the DESSA assessment was designed and standardized for ratings to be completed by educators and parents (or any adult living with the child at least par-time), as well as older students themselves. Furthermore, because bias can occur for any rater, we encourage different raters to compare and discuss ratings from a strength-based perspective. The DESSA is relatively unique in maintaining the same items and constructs across raters to enable direct comparison and discussion of individual items. Consistent ratings from multiple informants can help to affirm strengths and foster alignment on goals; inconsistent ratings can shed new insights

into student strengths. For example, a parent may rate their child lower in "cooperates with peers and siblings" while the student rates themselves relatively high on this item. Through discussion, the student may offer examples of their teamwork skills in out-of-school activities that the parent was not considering. Such discussions can help to obtain a richer, cross-informant and cross-environment understanding of the student and foster a sense of partnership based on equal contribution to the assessment that supports a collaborative, strength-based approach to decision making (LeBuffe & Shapiro, 2008).

Analytic Approaches. Analytic approaches involve the use of statistical procedures to help identify rater bias and evaluate its impact. One analytic approach is to look at whether two individuals who rate the same child make similar ratings. If independent raters show high agreement (i.e., inter-rater reliability), then bias tends to be less likely. Reliability studies of the DESSA-mini show good inter-rater reliability between two people within the same relationship to the child (e.g., teacher compared to teacher, parent compared to parent), with correlations that are relatively high (i.e., $r = .70-.81$) (Naglieri et al., 2011; 2014). Inter-rater reliability correlations for the full DESSA are similar (i.e., $r = .70-.84$) (LeBuffe et al., 2009, 2014). Differences do exist, but they are small in magnitude (i.e., less than 1 T-score point on average). Moreover, a medium sized correlation (i.e., $r = .42$) was found between

> **A rater <u>must</u> have sufficient exposure to the student over the four weeks prior to completing the DESSA.**

DESSA ratings from teachers and out-of-school staff, indicating good cross-informant agreement, a noteworthy finding across environments that can have different behavioral expectations (Shapiro et al., 2017).

A second analytic approach is to disaggregate the data and examine whether differences exist between subgroups of students. In a 3-year longitudinal study (Lee, 2022) of social and emotional competence growth using the DESSA-mini, demographic differences by gender, race/ethnicity, socioeconomic status (SES) were compared. At baseline, the magnitude of difference was small for gender and negligible for race/ethnicity. Over time, gender differences persisted, and males showed marginally lower rates of growth than females. Black students showed marginally lower rates of social and emotional competence growth than White students in Year 1, but not in subsequent years. SES differences were negligible at baseline and not associated with growth rates during the study. The differences observed in growth over time could be explained by rater bias, but also due to educational environments or interventions that worked differentially well in implementation for different groups of students. The study design was unable to distinguish between these possibilities.

A third analytic approach is to use a statistical technique, multilevel modeling, to gauge how much of the DESSA scores' variability is due to the individual differences among students vs.

other contextual factors (e.g., characteristics of the rater, classroom, school, program, or district). Research (Shapiro et al., 2016) shows that rater bias on DESSA-mini scores is quite low, accounting for only 16% of the variance in DESSA-mini scores (see also, Lee et al., 2018; Smith-Millman et al. 2017). This compares favorably to similar instruments. Moreover, rater bias can be reduced to only 10% of the variance in DESSA-mini scores once sources of variance that could be addressed in training and through implementation supports are taken into account (e.g., educator perceptions of SEL program implementation challenges and barriers to student learning) (Shapiro et al., 2016).

Implementation Practices. Implementation practices involve a series of recommended practices that schools, programs, and raters should follow prior to, and following, completing DESSA ratings that can help to reduce rater bias. First, because DESSA items refer to the frequency of specific, observable student behaviors, familiarity with the student being rated is essential. Therefore, DESSA ratings should only be made by raters who know the student well. The DESSA manuals (LeBuffe et al., 2009, 2014; LeBuffe et al., 2022a) specify that a rater <u>must</u> have sufficient exposure to the student over the four weeks prior to completing the DESSA.

Second, the DESSA manuals and training emphasize the importance of setting aside appropriate amounts of time for educators to complete the ratings so they can be made conscientiously, without undue fatigue, and do not compete with other work responsibilities. As noted above, an additional design safeguard is to carry out validity checks based on the time it takes raters to complete the DESSA. These checks identify raters that rushed through the assessment and may have provided unreliable responses.

Third, an educator's ratings may be, on occasion, atypically high or low for all their students, or for a specific subset of students. The DESSA reports permit educators to use interactive charting that allows ratings to be disaggregated according to

educational setting (classroom, school/program, and district) and student characteristics (gender, race/ethnicity, grade, free/reduced lunch, etc.) to pinpoint the nature of the outlying ratings. This information provides an opportunity for additional coaching to ensure all educators have the knowledge and resources they need to complete the assessment (e.g., training, time), and have adequate opportunities to reflect on and discuss bias. Such coaching need not be targeted; it can be provided in a supportive way that benefits all educators.

Finally, an additional practice to reduce bias is to engage educators in a discussion about their DESSA ratings with the rated students, parents, additional educators, and school staff (LeBuffe & Shapiro, 2008). Although these discussions have the explicit goal of building positive relationships, and discussing student strengths and needs, they also provide an impetus for seeing the human being behind each rating and may provide a form of accountability. That is, raters must be prepared to explain their ratings to key stakeholders. This accountability increases the likelihood that ratings will be made carefully and accurately.

Construct Bias

Construct bias, or cultural bias, occurs when the "construct" or social and emotional competency that we are measuring is not equally valued by, or relevant to, a particular culture or group of people. For example, some cultures may value competition and individual achievement, while others might emphasize collaboration and group achievement.

How do we Examine and Work to Reduce Cultural Bias?

The process to address construct, or cultural bias began with the initial item pool for the DESSA, as administered in English. All potential items were reviewed for mean score differences across select U.S. Census categories of race and ethnicity and items with large differences were eliminated (Shapiro & Lebuffe, 2006).

To further establish the cultural validity of the DESSA, Aperture Education has partnered with a firm specializing in multilingual communication and transcreation, and several cultural experts, to do a cultural appropriateness review of the DESSA constructs and items for administration in additional languages. This process involves three main steps:

1. Translate the DESSA into multiple languages including: Arabic, Bengali, Chinese, Haitian Creole, Hmong, French, Khmer, Korean, Portuguese, Russian, Spanish, Somali, Urdu, Vietnamese, to capture the nuances of language and the meaning of items.

2. Review of the DESSA constructs (i.e., the eight social and emotional competencies) and items by experts in education or psychology with native knowledge of the various cultural groups that speak the language in a North American context, noting any concerns about cultural misfit, and providing guidance to DESSA users to help interpret the results in a culturally sensitive way.

3. Engage diverse families and students in a discussion of the cultural relevance of the DESSA content, noting any remaining concerns, and modifying guidance to DESSA users, as appropriate.

These steps are designed to avoid items or constructs that might elicit different responses from various cultural groups or inadvertently advantage some cultural ways of being over others. Some cultural differences in the prevalence and meaning of specific DESSA items might still exist, as they would with any assessment. Therefore, as specified in the DESSA

manuals (LeBuffe et al., 2009, 2014; LeBuffe et al., 2022a), DESSA users should be sensitive to linguistic and cultural differences when interpreting the DESSA. Specifically, knowledge of the child's and family's culture will result in more sensitive interpretations of DESSA findings, and more useful recommendations to parents, educators, and young people.

Item Bias

Item bias occurs when students from different demographic, socioeconomic, or cultural groups *who have the same overall level of social and emotional competence* tend to be rated differently on the same DESSA item. For example, socially competent Native American students may tend to avoid eye contact with adults when speaking to them; socially competent European-American students may tend to look at adults when speaking to them. This *item* showed an obvious cultural bias (and was eliminated from the initial item pool of the DESSA). In a school district study using the final DESSA items, no statistically significant differences were detected between American Indian Alaska Native, other students of color, and white students (Chain et al., 2017).

How Do We Examine and Work to Reduce Item Bias?

In addition to the cultural review described above, the DESSA-High School Edition (LeBuffe et al., 2022a) and DESSA High School Edition-Student Self Report (LeBuffe et al., 2022b) also used a statistical technique called differential item functioning that examined whether different demographic groups responded differently to DESSA items. Small differences in social and emotional competence were found between genders, but differences based on race, ethnicity, and SES were negligible. Specifically, across racial/ethnic comparisons (i.e., white and Black, white and Hispanic, Black and non-Black, or Hispanic and non-Hispanic students), no DESSA item showed differences meeting the dual criterion of statistical significance and a non-negligible effect size.

> **All the DESSA assessment norms were developed using large, representative, national samples.**

DESSA reports also permit educators to utilize interactive charting that allows them to disaggregate data by student demographics and enables the exploration of potential bias in ratings at the student, classroom, school/program, and district levels.

Finally, DESSA results allow users to engage in individual item analysis. This process can elucidate the specific, item-level, strengths for each student. The approach can foster awareness that, even if a student's overall social and emotional competence may be low, they may have particular areas of strength. Such insights into student strengths may help to positively transform educator perceptions of their students and foster a more collaborative and hopeful partnership between schools, students, and families (LeBuffe et al., 2018).

Sample Bias

Assessment norms allow DESSA users to understand the relative standing of each student's score in relation to other students. Sample bias occurs when an assessment's norms are based on a specific, or non-representative group of students, and then applied to another group of students. For example, a test of academic achievement is normed on disproportionately high achieving school districts and then students from a typical school district are scored based on those norms.

How Does the DESSA Reduce Sample Bias?

Consistent with recommendations by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014), all the DESSA assessment norms were developed using large, representative, national samples. Specifically, the DESSA samples used to establish norms closely approximated the U.S. population in terms of gender, grade, geographic region, race, Hispanic ethnicity, and SES based on free or reduced-lunch eligibility status. The adequacy of the norms were independently reviewed (e.g., Atlas, 2010; Malcomb, 2010) and determined to be sufficiently large and diverse (Merrell & Gueldner, 2010).

National norms allow DESSA users to interpret, convey, and respond to the strengths and needs of the student, or groups of students, in relation to what can be expected of diverse students nationwide. National norms not only allow a broad perspective in the interpretation of individual student ratings, but also enables DESSA ratings generated within one administration environment to be compared to ratings obtained in other environments to understand the relative strengths and needs of students within various contexts. Although it is never appropriate to characterize the potential of a student or group of students by comparing them to other students, this reference point may help to allocate attention and resources to where they are most needed. Moreover, because national norms tend to change slowly, ratings can be compared across time to monitor progress and evaluate outcomes.

How Does Our Implementation and Training Address Bias?

Research (Shapiro et al., 2016) suggests professional development and reflection to foster culturally responsive and affirming practice may represent an important strategy to reduce rater bias. Therefore, Aperture Education provides professional development resources with the goal of supporting educators in providing valid, objective, and useful estimates of students' social and emotional competencies. This training is included as part of our implementation processes.

In addition to implementation practices to reduce rater bias described earlier, and intentional training provided on how to use and interpret the DESSA, Aperture Education offers all DESSA users access to grade-based Growth Strategies. These strategies are designed to support student development across all eight social and emotional competencies.

Finally, one study has observed a relationship between classroom teachers' own social and emotional competence and the growth in student DESSA scores (Lee et al., 2018). For this reason and others, Aperture Education has developed the Educator Social and Emotional Reflection and Training (EdSERT; Robitaille & LeBuffe, 2019)

professional development program to support the social and emotional knowledge, skill set, and well-being of educators. For instance, if a teacher is not aware of specific cultural differences, then they might misinterpret a behavior and rate a student low on a certain skill (e.g., a student who doesn't make eye contact when adults are speaking may be seen as having low relationship skills). EdSERT may help to reduce educator bias by increasing educators' understanding of SEL constructs. Specifically, EdSERT offers educators tools to reflect on their own social and emotional competencies, needs, and biases, and develop skills in areas such as empathy, perspective taking, respect, communication, authentic appreciation, and caring as they prepare to model and lead SEL activities with their students. The study by Lee and colleagues (2018) further suggests that school leaders who *proactively plan for implementation*, can moderate the impact of individual teacher social and emotional competence on the success of social and emotional learning initiatives.

Conclusion

The DESSA measures are brief, specific, strength-based behavioral rating assessment scales that incorporates several design characteristics which, in combination with statistical procedures and educator training, help to explore, identify, and reduce biases that can become embedded in student assessments. We cannot eliminate bias in all its forms, but we have a duty to work to minimize it, and the impact it can have on students. The field of SEL assessment is relatively new, and we plan to continue discovering new ways of identifying, understanding, and mitigating assessment bias, so that we can accurately measure and continuously improve the well-being of all students.

References

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

Atlas, J. A. (2010). Test review of the Devereux student strengths assessment. *The eighteenth mental measurements yearbook,* 178-180.

Chain, J., Shapiro, V. B., LeBuffe, P. A., & Bryson, A. M. (2017). Academic achievement of American Indian and Alaska Native students: Does social emotional competence reduce the impact of poverty. *American Indian and Alaska native mental health research (Online)*, *24(1)*, 1-29. http://doi.org/10.5820/aian.2401.2017.1

LeBuffe, P. A., & Shapiro, V. B. (2008). Supporting parent-professional collaboration through strength-based assessment. *Systems of Care*, 263.

LeBuffe, P. A., Shapiro, V. B., & Naglieri, J. A. (2009, 2014). *The Devereux Student Strengths Assessment (DESSA)*. Charlotte, NC: Apperson.

LeBuffe, P. A., Shapiro, V. B., & Robitaille, J. L. (2018). The Devereux Student Strengths Assessment (DESSA) comprehensive system: Screening, assessing, planning, and monitoring. *Journal of Applied Developmental Psychology*, *55*, 62-70. http://dx.doi.org/10.1016/j.appdev.2017.05.002

LeBuffe, P.A., Shapiro, V.B., Robitaille, J.L., & Naglieri, J.A. (2022a). *The Devereux Student Strengths Assessment High School Edition (DESSA-HSE)*. Aperture Education.

LeBuffe, P.A., Shapiro, V.B., Robitaille, J.L., & Naglieri, J.A. (2022b). *The Devereux Student Strengths Assessment High School Edition Student Self-Report (DESSA-HSE SSR)* [Manuscript in preparation]. Aperture Education.

Lee, J. (2022). Gender, racial-ethnic, and socioeconomic disparities in the development of social-emotional competence among elementary school students [Manuscript in preparation]. School of Social Welfare, University of California, Berkeley.

Lee, J., Shapiro, V. B., Kim, B. E., & Yoo, J. P. (2018). Multilevel structural equation modeling for social work researchers: An introduction and application to healthy youth development. *Journal of the Society for Social Work and Research*, *9*(4), 689-719. http://doi.org/10.1086/701526

Malcomb, K. K. (2010). Test review of the Devereux Student Strengths Assessment. *The eighteenth mental measurements yearbook,* 180-182.

Merrell, K. W., & Gueldner, B. A. (2010). *Social and emotional learning in the classroom: Promoting mental health and academic success.* New York: Guilford Publications.

Naglieri, J. A., LeBuffe, P. A., & Shapiro, V. B. (2011). Universal screening for social–emotional competencies: A study of the reliability and validity of the DESSA-mini. *Psychology in the Schools, 48*(7), 660-671. https://doi.org/10.1002/pits.20586

Naglieri, J. A., LeBuffe, P. A., & Shapiro, V. B. (2011, 2014). *The Devereux Student Strengths Assessment – Mini (DESSA-Mini): Assessment, technical manual, and user's guide.* Charlotte, NC: Apperson.

(References Cont.)

Robitaille, J.L. & LeBuffe, P.A. (2019). *The Educator Social and Emotional Reflection and Training (EdSERT) Program.* Fort Mill, SC: Aperture Education.

Rosas, S., Chaiken, L., & Case, J. (2006). *Research brief: Devereux Early Childhood Assessment ratings from parents and teachers in Delaware.* Retrieved December, 21, 2009 from https://centerforresilientchildren.org/wp-content/uploads/deca.pdf.pdf

Shapiro, V. B., Accomazzo, S., & Robitaille, J. L. (2017). In the same ballpark or a whole new ball game? Staff as raters of youth behavior. *Journal of Child and Family Studies, 26(4)*, 1051-1055. https://doi.org/10.1007/s10826-016-0632-1

Shapiro, V. B., Kim, B. K., Accomazzo, S., & Roscoe, J. N. (2016). Predictors of rater bias in the assessment of social-emotional competence. *International Journal of Emotional Education*, *8*(2), 25-44.

Shapiro, V. B., & LeBuffe, P. A. (2006). Using protective factors in practice: Lessons learned about resilience from a study of children aged five to thirteen. *Annals of the New York Academy of Sciences*, *1094(1)*, 350-353. https://doi.org/10.1196/annals.1376.048

Smith-Millman, M. K., Flaspohler, P. D., Maras, M. A., Splett, J. W., Warmbold, K., Dinnen, H., & Luebbe, A. (2017). Differences between teacher reports on universal risk assessments. *Advances in School Mental Health Promotion*, *10*(4), 235-249. https://doi.org/10.1080/1754730X.2017.1333914

[1]*Valerie Shapiro, co-author of the DESSA, assisted as an independent consultant not employed by Aperture Education on this piece.*

APERTURE EDUCATION

Aperture Education has empowered over 6,500 schools and out-of-school time programs across North America to measure, strengthen, and support social and emotional competence in K-12 youth and educators. The Aperture System includes the DESSA suite of strength-based assessments, CASEL™-aligned intervention strategies, and robust reporting, all in one easy-to-use digital platform. This system enables education leaders to make strategic, data-based decisions about SEL within their organizations. Aperture has supported more than one million students in their social and emotional growth and continues to develop innovative solutions to bring the whole child into focus. To learn more, visit www.ApertureEd.com.