



JONATHAN SCHWEIG, SUSAN BUSH-MECENAS, SY DOAN, MELISSA KAY DILIBERTI

Validity of Ratings from the Thrive for Teachers Evaluation System

PR-A1651-3

May 2023

Sponsored by the Detroit Public Schools Community District

About RAND

RAND is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. To learn more about RAND, visit www.rand.org.

Project Reports

RAND project reports convey information about a project to research sponsors. This document is intended for use by the Detroit Public Schools Community District, and further dissemination is at their discretion.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

© 2023 RAND Corporation

RAND® is a registered trademark.

About This Project Report

Beginning in 2019, the Detroit Public School Community District (DPSCD) developed the Thrive for Teachers and Thrive for School Administrators evaluation systems. These systems were first implemented in the 2020–2021 and 2021–2022 school years, respectively. The Thrive evaluation systems were developed in connection with DPSCD’s strategic plan, Blueprint 2020. Blueprint 2020, and its subsequent Blueprint 2021 extension plan, outlined the DPSCD’s mission, vision, core values, and strategic priorities. The Thrive evaluation systems, which were developed and designed in collaboration with teachers and school administrators, are an integral part of this Blueprint plan and align directly to two strategic priorities: (1) developing exceptional talent and (2) producing outstanding achievement results for students. The Thrive evaluation systems were designed to relate directly to DPSCD’s Vision of Excellent Instruction and leadership expectations for school administrators, such that the district’s expectations for excellence in teaching and administration would align with how employees are evaluated.

In accordance with the state of Michigan’s legislative statutes 380.1249 and 380.1249b, DPSCD requested assistance from the RAND Corporation to examine the technical qualities of the teacher and administrator ratings produced under the Thrive evaluation systems. Specifically, these laws state that external researchers must provide “an assurance that the adaptations or modifications [to the existing evaluation system] do not compromise the reliability, validity, or efficacy of the evaluation tool or the evaluation process.” This report details our investigation of Thrive evaluation system ratings during the 2021–2022 school year, and provides evidence of the reliability, validity, and efficacy of these systems.

This report is intended for the DPSCD staff who are responsible for maintaining the Thrive evaluation systems. This report’s objective is to document the current system implementation and the evidence brought to bear on the reliability, validity, and efficacy of the systems. Using the evidence collected, we synthesized key themes that emerged from the accumulated data. Consistent with Michigan legislative statutes, separate documents detailing these key themes, serving as an attestation to this work being conducted, will be produced and published.

RAND Education and Labor

This study was undertaken by RAND Education and Labor, a division of the RAND Corporation that conducts research on early childhood through postsecondary education programs, workforce development, and programs and policies affecting workers, entrepreneurship, and financial literacy and decisionmaking. The research reported here was sponsored by DPSCD. The opinions expressed are those of the authors and do not represent views of DPSCD.

More information about RAND can be found at www.rand.org. Questions about this report should be directed to jschweig@rand.org, and questions about RAND Education and Labor should be directed to educationandlabor@rand.org.

Acknowledgments

The authors are grateful to many current and former DPSCD administrators and staff for their support on this project, especially the following: Superintendent Nikolai Vitti; Senior Executive Director of Data, Assessment, and Accountability Mark Paik; Deputy Superintendent of Strategy and Academics Elizabeth Cutrona; Deputy Superintendent of Strategy and Academics Barbara Thompson (who is also former Deputy Executive Director, Assessment and Performance Management); and former Operations Associate Bradley Kingston.

This report benefited substantively from feedback from Chris Doss at the RAND Corporation and Jose Felipe Martinez at the University of California, Los Angeles. Maria Vega provided expert editing. We thank Rebecca Wolfe for her assistance with qualitative data collection. Any flaws or errors that remain are the sole responsibility of the authors.

Finally, we are indebted to the educators, school administrators and local superintendents who participated in the research, taking time to participate in our observation study and in focus groups with RAND.

Contents

About This Project Report	ii
What We Did.....	2
What We Found	3
References.....	5

(This page intentionally left blank.)

The Detroit Public Schools Community District (DPSCD) began developing the Thrive for Teachers evaluation system in 2019 and first implemented it in the 2020–2021 school year. The Thrive for Teachers evaluation system was designed in collaboration with educators such that DPSCD’s expectations for excellence in teaching would align with how teachers are evaluated. In accordance with the state of Michigan’s legislative statutes 380.1249 and 380.1249b, DPSCD requested assistance from the RAND Corporation to examine the technical qualities of the ratings produced under the Thrive for Teachers evaluation system. Specifically, these laws state that external researchers must provide “an assurance that the adaptations or modifications [to the

Components of the Thrive for Teachers Evaluation System

The system has four components that are weighted and combined together to form an overall score. Based on these overall scores, teachers receive one of four overall ratings: Ineffective, Minimally Effective, Effective, or Highly Effective.

- **Classroom Observations (45% of overall score for teachers with survey scores, 50% otherwise).** Teachers are observed and evaluated using a classroom observation rubric three times during an academic year. The Cycle 0 observation does not count toward their final evaluation score. The Cycle 1 and Cycle 2 observations occur in the second and third quarters, respectively, of the academic year. The observation rubric has four domains: Management, Culture, and Climate, Rigorous Content, Academic Ownership, and Demonstration of Learning.
- **Commitment to School Community (CtSC; 10% of overall score).** This rubric-based component evaluates a teacher’s support of local school initiatives, professional growth and collaboration, promotion of high academic and behavioral expectations, and partnership with students’ families. For this component, school administrators identify exemplary actions and practices that teachers can undertake to help cultivate the administrator’s vision of a positive school environment. With assistance from a district-provided guidebook, school administrators identify examples of evidence that show teachers are undertaking such efforts. Evidence can include observations of actions outside of classroom contexts, documents, and conversations among staff and administrators.
- **Student Survey (5% of overall score when available).** Students provide feedback on teaching and the quality of the learning environment on an annual survey questionnaire.
- **Student Growth (40% of overall score).** Teachers receive scores based on student growth calculated from two sources of student assessment data, with strong preference for state and district assessment data, and local assessment data if needed. Michigan state statutes require that 40 percent of a teacher’s total evaluation score be based on student growth.

existing evaluation system] do not compromise the reliability, validity, or efficacy of the evaluation tool or the evaluation process.”

Validation is an ongoing process: As more evidence accumulates across time, sources, or contexts, the balance of the evidence may shift. This aspect is particularly important for this study, which was conducted in the first full school year (2021–2022) school closures related to the coronavirus disease 2019 (COVID-19) pandemic. As is well established, the COVID-19 pandemic had tremendous effects on teachers and students and upended administrative processes in school systems, including in Detroit. For example, there was a mix of in-person and remote learning through winter 2022, state summative assessment programs were suspended in 2020, and federal accountability programs were suspended in both 2019–2020 and 2020–2021. By any definition, the 2021–2022 academic year was not a typical school year, which had implications for both the kinds of evidence and the quality of evidence that we were able to collect for this study.

What We Did

Determinations about reliability, validity, and efficacy are not yes-or-no propositions (Gehlbach and Hough, 2018; Kane 2006). Following best practices, we collected as much

Reliability, Validity, and Efficacy

Following best practices outlined in research (Kane, 2006; Bell et al., 2012; American Educational Research Association, American Psychological Association, and National Council of Measurement in Education, 2014), we define *reliability*, *validity*, and *efficacy* as follows:

- **Reliability** is the degree to which the Thrive for Teachers evaluation system produces scores that are consistent, precise, and relatively free of measurement error. Measurement error refers to random variation in evaluation scores that is not related to leadership quality. As one example of measurement error, by random chance, an observer might be inclined to be more severe or more lenient than other observers, and this aspect would introduce some random variation into observation scores that is not related to teaching effectiveness.
- **Validity** is the degree to which evidence and theory support interpretation of Thrive for Teachers Evaluation overall ratings for three proposed uses: (1) to identify areas of strength, as well as areas for growth and professional development; (2) to inform decisions around tenure; and (3) to guide internal decisions around the transfers of teachers between schools.
- **Efficacy** describes the capacity of the evaluation system to produce the intended results. For example, feedback processes from the system should provide teachers with information they can use to improve their practice.

evidence as possible during the 2021–2022 school year to illuminate both the strengths and the limitations of the Thrive for Teachers evaluation system and its components. Using all the evidence collected, we offer a professional judgment about the extent to which the strengths of the system outweigh its limitations, with a particular focus on whether the limitations of the system are likely to interfere substantially with the intended uses.

We examined all the evaluation data available for 3,375 DPSCD elementary, middle, and high school teachers from the 2021–2022 school year. Our examinations consisted of convening an expert panel to appraise the content validity of the observation rubric, looking at the score distributions for each Thrive for Teachers system component, inspecting relationships among components, and using linear regression to explore the extent to which component and overall system scores were associated with teacher and school characteristics. We also conducted a separate rater study to explore the accuracy and consistency of scores derived from the Classroom Observation rubric. In addition to these quantitative analyses, we also conducted interviews and focus groups with ten administrators and 30 teachers to learn more about implementation of the system and the extent to which educators considered the Thrive for Teachers evaluation system and its components accurate, trustworthy, and useful. Using the analyses of all the evidence we collected, we made the following determinations about the reliability, validity, and efficacy of the Thrive for Teachers evaluation system.

What We Found

Thrive for Teachers evaluation scores are intended to make claims about teacher effectiveness. All teachers are eligible for evaluation each year. Given the intended uses of these scores (to support professional development, inform tenure processes, inform teacher transfers), we did not find deficiencies in the system component scores or in the overall ratings that would critically interfere with these intended interpretations or applications. However, there are several specific issues with instruments and scoring rules that may compromise the potential of the system to improve teaching and learning for all students. We discuss the strengths of the Thrive for Teachers system and the specific issues, as follows.

- **Reliability:** The reliability coefficients we estimated are similar in value to those reported in studies analyzing state-approved instruments currently used in other teacher evaluation systems throughout the state of Michigan (including the Danielson Framework for Teaching and the Marzano Teacher Evaluation Model). Although the reliability of the system components is comparable to other state-approved tools, the extent of measurement error, particularly in the Observation and CtSC scores, suggests a lack of precision that could lead to substantial fluctuations from cycle to cycle or year to year. Although these fluctuations are most likely to occur between Effective and Highly Effective designations, some research suggests that these yo-yo effects in accountability outcomes can lead to fairness concerns and ultimately decrease trust in the system.
- **Validity:** A strength of the Thrive for Teachers evaluation system is that it was designed collaboratively with DPSCD leaders and teachers and is aligned to the district’s strategic

plan. There is strong evidence that the system is focused on important aspects of teaching effectiveness, and the individual system components are positively correlated—all of which supports the notion that valid inferences can be made about teaching effectiveness on the basis of Thrive scores. On the other hand, there are three specific weaknesses with the system that should be further investigated and addressed. First, the clustering of scores around the Effective and Highly Effective categories is a significant limitation of the system. In general, we would expect there to be a larger proportion than what we found of teachers that are Minimally Effective or Ineffective. Furthermore, the capacity of the scores to support professional development, inform tenure processes, and inform teacher transfers is diminished by the lack of variation in teacher ratings. Second, the way that scores are produced for the for the Student Survey component compresses a lot of underlying variation and decreases the utility of the survey-based scores. Finally, we found systematic differences in Thrive scores based on grade level and subject taught. Such differences do not reflect teaching effectiveness and can lead to fairness concerns and, ultimately, decrease trust in the system.

- **Efficacy:** We cannot speak directly to whether scores and effectiveness ratings improve teaching or student outcomes. Instead, we focus on whether teachers and administrators find the scores useful for this purpose. The teachers and administrators we spoke to believed that the Thrive for Teachers evaluation system focused on aspects of practice that were important for effective teaching. However, challenges arose in implementation across school sites. Teachers raised questions about whether evaluators could reasonably observe all elements of the rubric within each observation session. They also questioned whether the elements of the rubric would be equally evident in different classroom settings, particularly in special education settings or non-academic classes. Some teachers found the CtSC ratings unclear and potentially subject to evaluator bias; they identified variation in which activities “counted” as a contribution to school community and possible variation in both the expectations for different teachers and the accessibility of opportunities to participate in or lead specific, valued activities. Finally, teachers and administrators alike questioned the usefulness and accuracy of student surveys as part of their evaluation, given concerns over students’ ability to understand survey constructs and the usefulness of student satisfaction data to instructional practice improvement.

References

- American Educational Research Association, American Psychological Association, and National Council of Measurement in Education, *Standards for Educational and Psychological Testing*, 2014.
- Bell, C. A., D. H. Gitomer, D. F., McCaffrey, B. K. Hamre, R. C. Pianta, and Y. Qi, “An Argument Approach to Observation Protocol Validity,” *Educational Assessment*, Vol. 17, No. 2–3, 2012, pp. 62–87.
- Gehlbach, H., and H. J. Hough, *Measuring Social Emotional Learning through Student Surveys in the CORE Districts: A Pragmatic Approach to Validity and Reliability*, Policy Analysis for California Education (PACE), 2018.
- Kane, M. T., “Validity,” *Educational Measurement*, Vol. 4, 2006, pp. 17–64.