USING PSYCHOMETRIC ANALYSIS TO IMPROVE SOFT-SKILL ASSESSMENTS

SEPTEMBER 2022

EDITH YANG (MDRC)
PETER HALPIN
(UNC-CHAPEL HILL)
DANIEL HANDY (MDRC)

Postsecondary institutions, and particularly community colleges, are positioned well to train more and more of the nation's entry- and middle-skill workers.[1] They are increasingly recognizing the importance of equipping their students to master soft skills—the capabilities and habits that affect communication, social interactions, and problem-solving, also referred to as nonacademic skills, workplace skills, or essential skills—to match employer expectations in the labor market.[2] They are also looking for ways to increase the reliability and validity of measures used to assess such skill mastery.[3]

The New World of Work (NWoW), a program that promoted teaching and learning soft skills, was designed in 2012 and operated briefly in over 75 community colleges in California. The program consisted of a classroom component, a work-based learning component, and a credential-granting component. A companion brief details lessons from its implementation.[4] This brief focuses on the development and refinement of the credential-granting component. Specifically, it describes how psychometric analysis can inform the design of assessments used to grant credentials so that soft skills can be more accurately measured and employers and educators can recognize the credentials' value. Curriculum designers might find that psychometric analysis, in combination with other information (such as the perspectives of instructors, employers, and students), can help them determine how well their assessments capture the mastery of the soft skills students need to thrive in today's labor market.

mdrc
BUILDING KNOWLEDGE TO IMPROVE SOCIAL POLICY

PSYCHOMETRIC ANALYSIS EXPLAINED

Psychometrics concerns the reliability and validity of educational and psychological tests.[5] When students take a well-designed math exam, for example, the teacher can infer from students' scores which students have grasped the math concepts assessed and which students need more support. The teacher cannot observe the students' math abilities directly but can make inferences about their abilities from their test scores. Psychometric analysis can be used to test how well an assessment measures the level of various constructs, such as the measures for "achievements," "knowledge," or "skill mastery." In the case of NWoW, psychometric analysis was used as a preliminary test of each of its 10 soft-skills assessments to produce recommendations on improving the assessments for larger-scale use at more colleges.

THE NWOW ASSESSMENTS AND DIGITAL BADGES

Drawing on employer surveys, published research, and information from many panels of California employers, human resources managers, instructors, and students, the NWoW program-design team developed soft-skill assessments to measure the mastery of 10 soft skills they determined employers would value.[6] They identified these 10 skills using Mozilla Foundation's national comparison of College/Career Ready Competencies. Table 1 lists 3 of the 10 skills these assessments aimed to measure and the professional competencies associated with each skill. The three selected skills—adaptability, collaboration, and communication—are the examples illustrated throughout this document. NWoW instructors and students identified these three skills as being immediately applicable, relevant, and in demand in the twenty-first-century labor market.

The assessment for each of these skills is structured similarly, beginning with about 15 to 20 true/false, multiple choice, or matching questions, then proceeding to one video assessment and two free-response reflection questions. Students could be awarded digital badges on any of these assessments on two levels: They could earn academic badges for any of these skills by achieving a score of 70 percent on the assessment, with the instructor verifying their mastery of the specific soft skill. They could additionally earn employer-verified badges if the employer of their work-based learning experience verified their skill mastery by completing an employer verification form. The psychometric analysis described in this brief applies only to the academic badges.

MEASURING THE PERFORMANCE OF THE NWOW ASSESSMENTS

If NWoW's digital badges were to demonstrate to employers that they represented students' mastery of the 10 soft skills, the assessments leading to those badges needed to measure that mastery effectively. The NWoW team therefore partnered with the MDRC research team to examine and strengthen the soft-skills assessments used in NWoW's pilot phase. The research team examined the psychometric properties of the existing assessments as a first step to improving their performance in measuring the 10 skills, and to considering whether 10 distinct assessments were indeed necessary.

TABLE 1

NWOW SOFT-SKILL COMPETENCIES

SKILL	PROFESSIONAL COMPETENCIES
Adaptability	• Aware of and positively responds to change. • Has a flexible approach to work, which includes various work environments, roles, and tasks. • Takes into account diverse viewpoints and input to achieve work outcomes. • Handles stress, feedback, and setbacks with healthy coping mechanisms in order to learn from experience and continue to move forward.
Collaboration	• Builds and maintains mutually beneficial relationships by working with diverse groups or teams. This includes the use of technology tools to allow in-person and remote teamwork. • Incorporates a range of perspectives and cultural norms while reinforcing common ground and shared goals. • Applies a transformational leadership approach where one seeks input, incorporates feedback, implements new ideas, offers help, and engages all team members in order to promote shared responsibility. • Handles conflict constructively and views failure as an opportunity to learn.
Communication	• Presents information that is appropriate in content, professional in both tone and language, and tailored to the recipient/audience. • Uses digital media, social media, and other technology communication tools properly for work settings. • Understands basic etiquette and rules in non-verbal, verbal, and written communication to effectively and accurately convey meaning. • Uses attentive listening skills, which includes asking clarifying questions and summarizing information back to check for understanding.

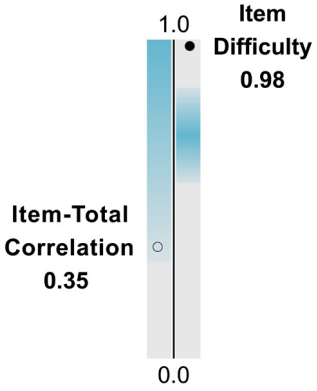
The research team ran descriptive statistics and conducted preliminary analyses using item response theory, a type of psychometric analysis that assesses both how well an overall assessment measures the mastery of an intended skill or construct and how well each item within an assessment contributes to the overall measure.[7] The analysis examines psychometric properties such as item difficulty, item-total correlation, total score distributions, total score correlations, Cronbach's alpha, and likelihood ratio tests. These terms are defined and discussed in the assessment examples provided later.

To conduct the analysis, the research team used assessment scores and item responses (from which individual identification data had been removed) from NWoW assessments that were administered from 2016 through 2018 across 16 postsecondary institutions in California. The data set had an average of 3,000 student respondents per assessment. Using the assessment data for the soft skills "adaptability," "collaboration," and "communication," this section illustrates how psychometric concepts can be applied.

Figure 1 shows examples of assessment items and their values for item difficulty and item-total correlation. Item difficulty is a measure of how difficult an item is, determined by how many

FIGURE 1

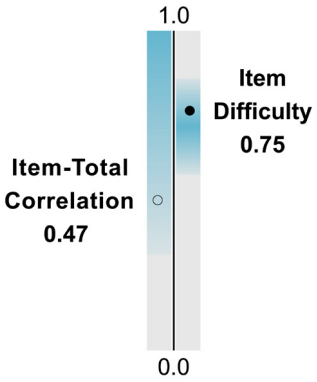
ITEM-LEVEL MEASURES: ITEM DIFFICULTY AND ITEM-TOTAL CORRELATION



ADAPTABILITY ASSESSMENT, ITEM 16:

It was your first performance evaluation and you were really disappointed that there was only one rating of 'exceeds expectations', most were for 'meets expectations', and there was even one rating of 'needs to improve'. What would you do next?

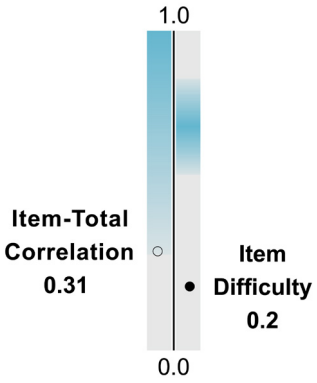
a. You decide you might as well start looking for a new job that doesn't require the one you need to improve.
b. You review the criteria and the comments and come up with a plan to improve the one where you were rated as needing to improve.
c. You talk to some of your friends at work about how unfair you thought your performance evaluation was.
d. You talk to some of your friends at work about how unfair you thought your performance evaluation was.



COLLABORATION ASSESSMENT, ITEM 9:

Put the team building stage description in the order in which they normally occur.

a. Team members may be anxious and cautious as they get to know one another. (1)
b. Team members trust and accept one another and perform as a cohesive unit. (4)
c. Team members become more positive about the team as a whole. (3)
d. Team members may question team goals and resist taking on tasks. (2)



COMMUNICATION ASSESSMENT, ITEM 9:

You see yourself as a really nice person. You are open and always willing to listen. Your communication style would best be described as:

a. Process focused
b. Action focused
c. People focused

SOURCE: Calculations from NWoW pilot assessment data. Correct responses are depicted with blue text.

NOTE: There is not a hard and fast rule about the acceptable values of item difficulty and item-total correlation. The blue shading indicates ranges where most of the items in each assessment should fall. Generally, item-total correlation values closer to 1 and item difficulty values around the passing score (0.7 in the case of NWoW assessments) are desirable.

respondents provided the correct answer. For items worth one point (correct or incorrect, or "binary" items), item difficulty is determined by the proportion of respondents who answered the item correctly. For items worth more than one point, item difficulty is determined by the mean score—that is, the number of points received as a proportion of total points available. More difficult items have lower values of item difficulty; easier items have higher values of item difficulty.[8] Item-total correlation reflects how strongly each item is related to the soft-skill construct that is being measured. It is calculated by correlating the score of each item on the assessment with the total score on the assessment without the item. Items are deemed better when they are strongly correlated with the total score. The field of psychometrics generally considers item-total correlations around 0.4 or higher to be reliable.[9]

From these two measures, assessment developers can more easily identify items that may be too easy or that have low correlation with the other items in the assessment. For example, the sixteenth item on the adaptability assessment is very easy: 98 percent of students chose the correct response. The incorrect response choices (in black) include "social desirability cues" that are clearly negative. In psychology, the term social desirability refers to people's tendency to answer a question in a way that they think is more socially acceptable so that they are viewed more favorably.[10] In this example, there are answer choices that suggest complaining to others and quitting one's job for another that does not encourage self-improvement. It is unlikely that respondents will choose these responses, since these would be generally perceived as negative traits. As a result, students may choose the correct response even without knowing anything about adaptability. The item also has low item-total correlation (0.35), suggesting that the question may not be as well correlated with the adaptability construct as other items on the assessment. One way to increase the difficulty of this item would be to replace the response options in black with options that are more neutral in social desirability but that do not capture adaptability as well as the correct response in blue.

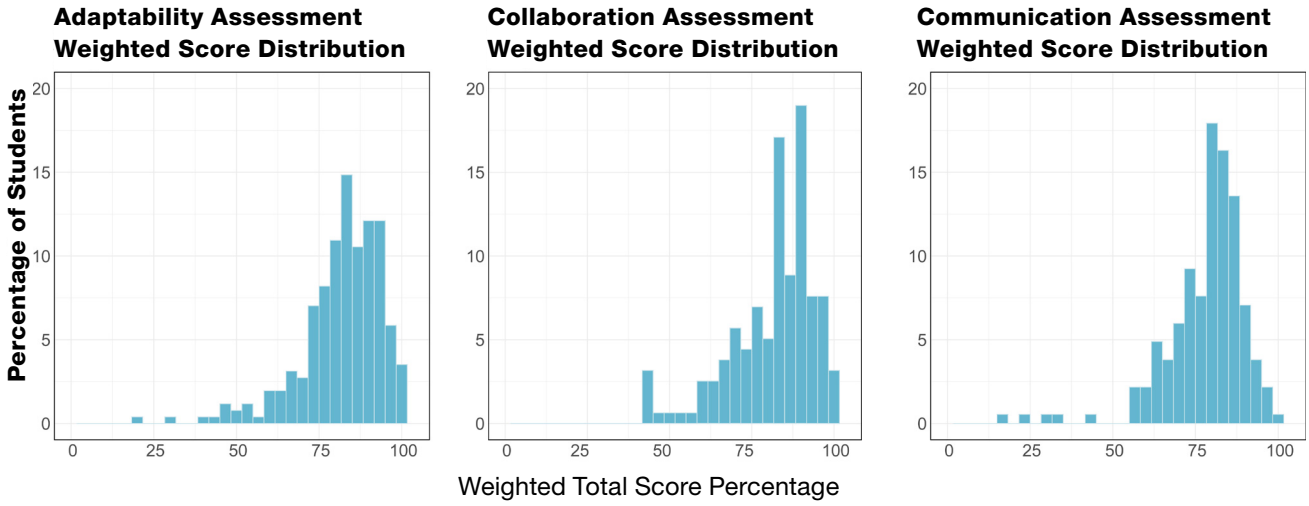
The second example, item 9 from the collaboration assessment, shows a question that is moderately difficult (with 75 percent of respondents answering correctly) and whose correlation with the other items in the assessment is acceptably strong (0.47). The question is about team building (clearly important for effective collaboration) and asks students to put specific stages in order. If assessment developers wanted to make this item more challenging, they could add more response options that represent intermediate stages of team building. Alternatively, since this item is more strongly correlated with the overall assessment construct than the other items in the assessment, developers could choose to weight the question more, or to give partial credit if a respondent puts some but not all the response choices in the correct sequence.

Finally, item 9 on the communication assessment is both extremely difficult (with only 20 percent of respondents answering correctly) and weakly correlated with the other items in the assessment (0.31). This item should probably be eliminated since the item-level metrics suggest that it does not measure the communication construct well.

In addition to item-level measures, psychometric analysis can help assessment developers examine how each assessment performs as a whole and compared with other assessments. Figure 2 shows the distribution of total scores on the adaptability, collaboration, and communication assessments. These distributions are helpful for at-a-glance perspectives of whether the overall assessments are too easy or too difficult.

FIGURE 2

ASSESSMENT-LEVEL STATISTICS AND CROSS-ASSESSMENT PROPERTIES



CRONBACH'S ALPHA

Adaptability	Collaboration	Communication
0.68	0.60	0.58

SOURCE: Calculations using NWoW assessment data from the pilot phase.

The horizontal axes show the proportion of total points that students achieved, and the vertical axes show the percentage of students who received each score. The total score distributions show that for each of the three assessments, most students achieved a score of 75 percent or higher, and very few students scored below 50 percent. The main concepts designers should focus on in these graphs are the skew of a distribution (whether there are more values concentrated on either the right or the left side of a graph), and any ceiling or floor effects (whether there are many outliers, or large numbers of extremely high or extremely low scores). These distributions have a slight negative skew (more values with higher scores on the right side of the distribution) and no apparent ceiling effects. To test whether the different assessments were measuring different things or were redundant to some degree, the research team also calculated how closely the total scores across all assessments were correlated with each other (not shown). These descriptive statistics did not raise concerns about score distributions or redundancy that assessment designers would need to address.

Like the item-total correlation estimate, Cronbach's alpha is a measure of internal consistency that is commonly used to test whether items in a scale are closely enough related to each other.[11] It is a lower-bound estimate on the overall reliability of a total score. Alpha will be large when all the items on an assessment are strongly correlated with each other. When some items are not strongly correlated, alpha will decrease. Minimum acceptable levels of reliability range from 0.6 to 0.8.[12] The three assessments appear to be adequately reliable, with the communication assessment being slightly less reliable than desired. For that assessment, developers might consider adding some more relevant questions while removing other items with weak item-total correlation.

Each assessment is intended to measure only a single construct. For example, the collaboration assessment should measure only collaboration. It should not assess both collaboration and communication. If an assessment measures more than one construct, its conceptual framework or the assessment items require further refinement. One way to evaluate whether an assessment measures more than one construct is to use a likelihood ratio (LR) test that compares a single-factor item response theory (IRT) model (which predicts mastery of one skill) with a two-factor IRT model (which predicts mastery of more than one skill). The "factors" are the soft-skill constructs measured by an assessment, and the results of the likelihood ratio test show which model better fits the data. If the two-factor model does not fit the data substantially better than the single-factor model, it means that a single factor or construct is sufficient to explain the data.

Table 2 shows the LR tests for the adaptability, collaboration, and communication assessments. The first two columns show the likelihoods of the data fitting the two models. The log is used to rescale the likelihoods of model fit so that tests of statistical significance can be conducted using the chi-square statistic. For interpretation, the most important statistics in the table are the p-values in the rightmost column. The single-factor model is considered sufficient if the p-value of the LR test is larger than 0.05. A p-value that is lower than 0.05 means that it is more than 95 percent likely that the assessment measures two soft skill constructs and not one.[13] Since the p-values for the adaptability and communication assessment LR tests are much higher than 0.05, these assessments are well explained by a single factor. The collaboration assessment has a p-value of 0.001, which suggests that it may be measuring more than one construct and should be further investigated. For example, there may have been only a few unreliable items on the assessment that have affected the model fit. Another possibility is that the collaboration construct may be more meaningful if it is broken into two or more additional constructs.

TABLE 2

LIKELIHOOD RATIO (LR) TESTS OF ADAPTABILITY, COLLABORATION, AND COMMUNICATION CONSTRUCTS

CONSTRUCT	LOG LR, ONE-FACTOR IRT	LOG LR, TWO-FACTOR IRT	CHI-SQUARE	DEGREES OF FREEDOM	P-VALUE
Adaptability	-2,123.55	-2,116.47	14.16	20	0.822
Collaboration	-1,511.46	-1,488.41	46.09	21	0.001
Communication	-1,341.29	-1,330.52	21.54	18	0.253

SOURCE: Calculations using NWoW assessment data from the pilot phase.

LESSONS FROM THE PSYCHOMETRIC ANALYSIS

Findings from psychometric analyses can be used to improve the quality of assessments so that they more accurately measure the soft skills they are intended to measure. First, the findings pinpoint easy questions that may not contribute much to measuring how much a person has learned or mastered. Item difficulty combined with item-total correlation can help assessment developers consider removing or revising questions that may not measure a skill in the most effective way. A closer look

at the results of the preliminary psychometric analyses on the NWoW assessments revealed that many of the easy items and the items with low item-total correlation had answer choices that were clearly incorrect or could be easily eliminated because of social desirability cues, rather than what learners would have picked up from a class or training program.

Of course, there are considerations in designing assessments that preclude eliminating all easy items or eliminating all items that have low correlation with the remainder of the assessment. Only including difficult items may discourage engagement or persistence. Additionally, it may be desirable to also observe whether the respondents have become better at reading and recognizing social cues after having used the curriculum. High-quality assessments, then, include a balance of items. Looking at the distribution of total scores across student test takers can give assessment designers an at-a-glance view of how well the assessments test mastery of certain constructs.

NWoW designers were aiming for a passing score of 70 percent for each assessment to indicate comprehension or mastery of a soft skill. However, across all the assessments, most students scored above 75 percent, so having a uniform passing score across all assessments may not be particularly meaningful or represent mastery across all assessments. One way to align the passing score with skill mastery is to remove items that are easy or have low item-total correlation. Another is to weight the items so that those that are more highly correlated with each other are given more points.

Finally, fitting the assessment scores to IRT models can help assessment developers determine whether each assessment is testing a distinct soft skill. If LR tests show that an assessment is probably measuring more than one construct, it might be time to revisit whether all 10 NWoW skills are in fact distinct, or whether specific items on that assessment are testing for mastery of a different skill.

NEXT STEPS IN IMPROVING ASSESSMENT PERFORMANCE

Conducting preliminary psychometric analysis on a pilot version of a soft-skills curriculum can be an opportunity to identify some areas for refinement while preparing for a larger rollout of a program. The preliminary analyses conducted on the NWoW assessments allowed the team to identify questions that might have been too easy and that might not have provided a meaningful measure of skill mastery, questions that may have contained response choices that measured proficiency in social desirability cues rather than skills, and assessments where the combination of items did not necessarily measure the desired construct.

Had NWoW continued to operate, the assessment developers might have used these results to revise some of the items in these assessments. Improving the performance of assessments is an iterative process, and additional psychometric analyses on those revised assessments might then have provided better information on how to refine the assessments further.

In addition to understanding how the assessments performed separately, it would also be important to understand how NWoW's 10 soft skills were related to each other. Confirmatory factor

analysis—another IRT modeling approach that allows researchers to model multiple constructs simultaneously—could have addressed how distinct the 10 NWoW skills were from each other.[14] The approach could have yielded answers to additional questions including:

- Are all 10 of the soft skill constructs different from each other, or are some of them highly correlated with one another?
- Is there evidence of a smaller number of "higher-order" social skills that explain the associations among the 10 skills measured by NWoW?

Finally, it could be relevant to produce more reliable test scores by using IRT rather than simple total scores for use in an evaluation. That is, the IRT models tested in Table 2 could also be used to produce scale scores that more precisely measure skill mastery, which could lead to improved precision in subsequent statistical analyses of program performance.

As with any type of quantitative analysis, the findings from the psychometric analysis provide only part of the information designers need to improve how well assessments measure mastery of distinct skills. First, the assessment data were limited in that they were not linked to student background characteristics (such as race and ethnicity, age, parental status, or English proficiency status) that might be correlated with how students understood or performed on the assessments. The assessments might therefore predict skill mastery better for some groups of students than others.

Combining the analyses with other data sources and information is also important, since assessments such as these do not speak to on-the-job skill application or performance. Soft-skill applications may also look different in industries or sectors—for example, communication in an engineering job might be different from communication in a job in health care.[15] In NWoW, the technical assistance team recommended that program developers and instructors couple the assessment results with employers' perspectives on student performance to better measure student progress and learning. The validity of the assessments could have been further explored by talking with students who took the NWoW course and NWoW's employer partners about their experiences with and perspectives of the NWoW program. This broader approach could have brought the assessment results closer to their application in real-world work experiences.

NOTES AND REFERENCES

1 Caitlin Dewey, "States Look to Community Colleges to Fill Labor Gap," *Stateline* (https://www.pewtrusts.org/en/research-and-analysis/blogs/stateline/2022/04/14/states-look-to-community-colleges-to-fill-labor-gap, 2022).

2 Robert I. Lerman, "Are Employability Skills Learned in U.S. Youth Education and Training Programs?" *IZA Journal of Labor Policy* 2, 6 (2013).

3 Martin R. West, Matthew A. Kraft, Amy S. Finn, Rebecca E. Martin, Angela L. Duckworth, Christopher F. O. Gabrieli, and John D. E. Gabrieli, "Promise and Paradox: Measuring Students' Non-Cognitive Skills and the Impact of Schooling," *Educational Evaluation and Policy Analysis* 38, 1 (2016): 148–170.

4 Hannah Dalporto and Marco Lepe, "Implementing Soft-Skills Programs in a Postsecondary Setting: Lessons from the New World of Work" (New York: MDRC, 2022).

5 Joint Committee for Educational and Psychological Testing, *Standards for Educational and Psychological Testing* (Washington, DC: American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014).

6 The 10 soft skills for which NWoW developed assessments are adaptability, analysis solution mindset (which refers to problem-solving skills), collaboration, communication, digital fluency, empathy, entrepreneurial mindset (which refers to the drive to develop new products), resilience, self-awareness, and social diversity awareness (which refers to sensitivity to differences in backgrounds and beliefs).

7 Frances M. Yang and Solon T. Kao, "Item Response Theory for Measurement Validity," *Shanghai Archives of Psychiatry* 26, 3 (2014): 171–177.

8 Richard J. McCowan and Sheila C. McCowan, *Item Analysis for Criterion-Referenced Tests* (Buffalo, NY: Buffalo State College, Center for Development of Human Services, 1999).

9 Andrew L. Comrey and Howard B. Lee, *A First Course in Factor Analysis*, 2nd edition (Hillsdale, NJ: Lawrence Erlbaum Associates, 1992).

10 Mario Callegaro, "Social Desirability," pages 826–826 in Paul J. Lavrakas (ed.), *Encyclopedia of Survey Research Methods* (Thousand Oaks, CA: Sage Publications, 2008).

11 B.S. Everitt and A. Skrondal, *The Cambridge Dictionary of Statistics*, 4th edition (New York: Cambridge University Press, 2010).

12 U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, and What Works Clearinghouse, *What Works Clearinghouse Procedures and Standards Handbook*, Version 5.0 (Washington, DC: U.S. Department of Education, 2022).

13 The value of 0.05 is often used to determine statistical significance, although other conventional values include 0.01 and 0.001.

14 Everitt and Skrondal (2010).

15 Jason A. Tyszko and Robert G. Sheets, *Co-Designing Assessment and Learning: Rethinking Employer Engagement in a Changing World*, Occasional Paper No. 39 (Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment, 2019).

ACKNOWLEDGMENTS

This brief is dedicated to the memory of Chase Johnson, whose invaluable efforts on this study showcased his passion for improving higher education and helping those in need. Thanks also to the California Community Colleges Chancellor's Office and the Foundation for California Community Colleges for their partnership and support throughout the study. The authors would also like to thank Rajinder Gill from the New World of Work and Hannah Dalporto, Marco Lepe, Cynthia Miller, Rosario Torres, Mary Visher, and Evan Weissman from MDRC for their significant research and management contributions to the project. Additionally, we appreciate Marjorie Dorimé-Williams, DeShawn Preston, and Sue Scrivener for reviewing and providing valuable insights on the brief, Audrey Yu for data analysis and visualization expertise, and Parker Cellura for coordinating of this brief so carefully. Finally, Joshua Malbin edited the report and Ann Kottner prepared it for publication. Funding for the New World of Work project and this brief was provided by the Institute of Education Sciences, U.S. Department of Education, through grant R305A170304 to MDRC. Opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Dissemination of MDRC publications is supported by the following organizations and individuals that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: The Annie E. Casey Foundation, Arnold Ventures, Charles and Lynn Schusterman Family Foundation, The Edna McConnell Clark Foundation, Ford Foundation, The George Gund Foundation, Daniel and Corinne Goldman, The Harry and Jeanette Weinberg Foundation, Inc., The JPB Foundation, The Joyce Foundation, The Kresge Foundation, and Sandler Foundation.

In addition, earnings from the MDRC Endowment help sustain our dissemination efforts. Contributors to the MDRC Endowment include Alcoa Foundation, The Ambrose Monell Foundation, Anheuser-Busch Foundation, Bristol-Myers Squibb Foundation, Charles Stewart Mott Foundation, Ford Foundation, The George Gund Foundation, The Grable Foundation, The Lizabeth and Frank Newman Charitable Foundation, The New York Times Company Foundation, Jan Nicholson, Paul H. O'Neill Charitable Foundation, John S. Reed, Sandler Foundation, and The Stupski Family Fund, as well as other individual contributors.

The findings and conclusions in this report do not necessarily represent the official positions or policies of the funders.

For information about MDRC and copies of our publications, see our website: www.mdrc.org.

Copyright © 2022 by MDRC®. All rights reserved.

NEW YORK
200 Vesey Street, 23rd Flr., New York, NY 10281
Tel: 212 532 3200

OAKLAND
475 14th Street, Suite 750, Oakland, CA 94612
Tel: 510 663 6372

WASHINGTON, DC
750 17th Street, NW, Suite 501
Washington, DC 20006

LOS ANGELES
11965 Venice Boulevard, Suite 402
Los Angeles, CA 90066