# Balance the Equation

## Student Survey Psychometric Analysis: Executive Summary

This summary describes the psychometric analysis conducted on the student survey Cohort 2 data from the Gates Foundation's Balance the Equation Algebra I Grand Challenge project. The project aimed to identify and study innovative math programs that held promise for improving outcomes for Black and Latino students, and all students experiencing poverty. All studies were designed to investigate the impact of the programs on improving students' experience in math class and math identity, using a common survey across studies. The goal of these analyses was to examine evidence of validity and reliability of the survey scores to support conclusions regarding the constructs identified and to make comparisons over time (i.e., baseline and outcome) and across groups (i.e., treatment and control groups).

### Data and Instrumentation

We analyzed survey data from 1,538 students who responded to the survey at baseline data collection (fall 2022) and 1,611 students who responded at outcome data collection (spring 2023). The survey collected responses for three constructs associated with class environment, including sense of belonging in math class, engagement in math class, enjoyment in math class, and two constructs associated with math identity, including sense of being a "math person" and perception of math as useful for reaching goals (Exhibit 1). Most of the items were chosen from existing instruments with evidence of validity and reliability, while adapting the language of some items to the specific context of this study. In some cases, the American Institutes for Research® (AIR®) team combined items from different scales into one single scale. The AIR team created additional items to complement the existing scales. For example, the *sense of belonging in math class* scale was created with nine items from three scales proposed by Maloney and Matthews (2020), three items from one scale by Hanson (2017), and three new items (Exhibit 1).

### Analytic Procedure

We conducted a series of confirmatory factor analysis (CFA) separately for baseline data and outcome data to examine whether the items represent the five constructs theoretically conceptualized at each data collection. CFA results were evaluated through the chi-square for model fit, and commonly used fit indices, such as the comparative fit index (CFI), the Tucker-Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR). Models that did not achieve adequate fit were modified based on the results (i.e., factor loadings, modification indices) and, guided by the structure of the original scales as published by their authors, their fit was assessed again. Once the final scales were determined, we examined the internal consistency of the final scales using Cronbach's alpha. Definitions of these analyses and thresholds for adequate fit are presented in the notes of each results table presented below.

We also compared the stability of the psychometric properties over time (baseline and outcome) and across groups (treatment and control groups). One important assumption made when using an instrument to compare students' scores over time and across groups is that the item parameters do not depend on the timing of the administration or students' characteristics. Violations of this assumption imply that differences across groups or over time could not be uniquely attributed to students' true standing on the constructs of interest, but would be conflated with changes in the instrument itself. To test this assumption, we conducted measurement invariance analyses (Widaman & Olivera-Aguilar, 2023) over time (baseline and outcome) and across groups (treatment and control). Measurement invariance analysis consists of fitting a series of nested models with increasing constraints in the item parameters: first, equality constraints across groups or over time are tested in the item loadings and then in the item intercepts. Models with equality constraints in the loadings and intercepts with good fit indicate that any differences observed between groups or over time represent true differences in the construct and not in the way the instrument functions.

**Finding 1: We found validity evidence based on internal structure for reporting nine scales.** The CFA results do not support the use of five scale scores as proposed. Instead, the results support nine scales, closely aligned with the original scales as proposed by their authors: (1) belonging: support from math teacher, (2) belonging: support from math class, (3) sense of belonging in math class, (4) behavioral engagement, (5) cognitive engagement, (6) enjoyment in math class, (7) identity as a "math person," (8) others see me as a math person, and (9) math relevance (see Exhibit 2 for a comparison of the proposed and new scales and Exhibit 3 for model-fit information). The final scales with their respective items and item loadings are shown in Exhibit 4.

**Finding 2: All scales yield reliable scores.** All nine scales show Cronbach's alpha reliability values ranging from 0.69 to 0.91 (Exhibit 2). These results indicate that the scales yield scores with high internal consistency reliability.

**Finding 3: We found evidence to support the use of scores in comparisons across groups and over time.** We found evidence of the stability of the item loadings and intercepts over time (Exhibit 5) and across groups (Exhibit 6) across all new scales. These results indicate that the new scales can be used to evaluate changes from baseline to outcome, and differences between treatment and control groups.

## CONCLUSION

The analysis conducted provides validity and reliability evidence for nine scale scores closely aligned with the original sources. The results indicate evidence of score stability that supports the use of the scores for comparisons over time (i.e., baseline and outcome) and across groups (i.e., treatment and control groups).

# Exhibits

### Exhibit 1. Proposed Survey Constructs

| Outcome | Construct | Number of Items | Source | Original Scale |
|---|---|---|---|---|
| Experience in math class | Sense of belonging in math class | 9 | Maloney, T., & Matthews, J. S. (2020). Teacher care and students' sense of connectedness in the urban mathematics classroom. *Journal for Research in Mathematics Education, 51*(4), 399–432. | The nine items from Maloney and Matthews (2020) scale were distributed in three scales: *emotional support from teacher*, *my math class is like a family*, and *my contributions are valued*. |
| | | 3 | Hanson, J. (2017). Determination and validation of the Project for Educational Research That Scales (PERTS) survey factor structure. *Journal of Educational Issues, 3*(1), 64–82. | Sense of belonging |
| | | 3 | New items | |
| | Engagement in math class | 13 | Fredricks, J. A., Wang, M. T., Schall, J., Hofkens, T. L., & Parr, A. (2016). Using qualitative methods to develop a survey measure of math and science engagement. *Learning and Instruction, 43*, 5–15. | The 13 items from Fredricks et al. (2016) were distributed in two scales: behavioral engagement and cognitive engagement. |
| | Enjoyment in math class | 11 | Fredricks, J. A., Wang, M. T., Schall, J., Hofkens, T. L., & Parr, A. (2016). Using qualitative methods to develop a survey measure of math and science engagement. *Learning and Instruction, 43*, 5–15. | Emotional engagement |
| | | 1 | New item | |
| Math identity | Sense of self as a "math person" | 4 | OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* OECD Publishing. | Mathematics self-concept |
| | | 2 | Cribbs, J. D., Hazari, Z., Sonnert, G., & Sadler, P. M. (2015). Establishing an explanatory model for mathematics identity. *Child Development, 86*(4), 1048–1062. | Recognition |
| | | 1 | New item | |
| | Perception of math as useful for reaching goals | 5 | OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* OECD Publishing. | Intrinsic and instrumental motivation for mathematics |
| | | 3 | New items | |

### Exhibit 2. New Survey Constructs

| Proposed Construct Scale | Original Scale | New Scale | Number of Items | Reliability Baseline | Reliability Outcome |
|---|---|---|---|---|---|
| Sense of belonging in math class | *Emotional support from teacher,* and *my contributions are valued* (Maloney & Matthews, 2020) | Belonging: support from math teacher | 5 | .83 | .84 |
| | *My math class is like a family* (Maloney & Matthews, 2020) | Belonging: support from math class | 3 | .69 | .74 |
| | Sense of belonging (Hanson, 2017)<br>One new item | Belonging: sense of belonging in math class | 4 | .88 | .90 |
| Engagement in math class | Behavioral engagement (Fredericks et al., 2016) | Engagement: behavioral engagement | 5 | .83 | .84 |
| | Cognitive engagement (Fredericks et al., 2016) | Engagement: cognitive engagement | 6 | .86 | .87 |
| Enjoyment in math class | Emotional engagement (Fredericks et al., 2016)<br>One new item | Enjoyment in math class | 6 | .87 | .88 |
| Math identity | Mathematics self-concept (OECD, 2013) | Identity: math identity as a “math person” | 4 | .79 | .79 |
| | Recognition (Cribbs et al., 2015)<br>New item | Identity: others see me as math person | 3 | .71 | .73 |
| Math relevance | Intrinsic and instrumental motivation for mathematics (OECD, 2013) | Math relevance | 4 | .90 | .91 |

## Exhibit 3. Confirmatory Factor Analysis Model Fit Results

| Model | Chi-Square (df) | RMSEA (95% C.I.) | CFI | TLI | SRMR |
|---|---|---|---|---|---|
| **Belonging (3-factors baseline)** | 249.23 (51) | .050 (.044, .057) | .973 | .966 | .028 |
| **Belonging (3-factors outcome)** | 373.50 (51) | .063 (.057, .069) | .967 | .957 | .030 |
| **Engagement (2-factors baseline)** | 151.75 (34) | .048 (.041, .056) | .983 | .978 | .020 |
| **Engagement (2-factors outcome)** | 163.73 (34) | .049 (.042, .057) | .984 | .978 | .022 |
| **Enjoyment (1-factor baseline)** | 47.37 (9) | .055 (.040, .070) | .938 | .909 | .017 |
| **Enjoyment (1-factor outcome)** | 99.86 (9) | .081 (.067, .096) | .980 | .967 | .022 |
| **Math identity (2-factors baseline)** | 125.91 (13) | .078 (.066, .090) | .965 | .943 | .034 |
| **Math identity (2-factors outcome)** | 111.98 (13) | .070 (.059, .083) | .971 | .954 | .031 |
| **Math relevance (1-factor baseline)** | 14.96 (2) | .068 (.039, .102) | .996 | .988 | .009 |
| **Math relevance (1-factor outcome)** | 48.81 (2) | .124 (.095, .156) | .989 | .967 | .014 |

*Note.* df = degrees of freedom; C.I. = confidence interval; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root mean residual. A model was considered to have an adequate fit to the data when the values of the fit indices were CFI and TLI > .90, and RMSEA and SRMR < .08, while models with values of CFI and TLI > .95, and RMSEA and SRMR < .05 were considered to have a good fit to the data (see West et al., 2012, for various proposed cutoff values).

## Exhibit 4. Final Scales and Item Loadings

| Model | Item | Baseline | Outcome |
|---|---|---|---|
| **Belonging: support from math teacher** | I feel that I can talk to my math teacher about things that are bothering me.[1] | .691 | 0.759 |
| | My math teacher is someone I can count on to help me.[1] | .741 | 0.804 |
| | I believe that my math teacher is interested in the things I have to say.[1] | .750 | 0.751 |
| | My math teacher knows that I can do good work.[1] | .701 | 0.653 |
| | I have a lot of opportunities to get involved in my math class.[1] | .657 | 0.659 |
| **Belonging: support from math class** | My math class is like a family.[1] | .645 | 0.728 |
| | Students in math class help each other, even if they are not close friends.[1] | .646 | 0.667 |
| | When I do well in math, the other students in class are usually happy for me.[1] | .665 | 0.685 |
| **Belonging: sense of belonging** | I feel accepted in this class. | .842 | 0.865 |
| | I feel like I belong in this class.[2] | .850 | 0.871 |
| | I feel like I can be myself in this class.[2] | .782 | 0.767 |
| | I feel comfortable in this class.[2] | .765 | 0.817 |
| **Engagement: behavioral engagement** | I stay focused.[3] | .692 | 0.725 |
| | I answer questions in math class.[3] | .659 | 0.69 |
| | I put effort into learning math.[3] | .803 | 0.816 |

| Model | Item | Baseline | Outcome |
|---|---|---|---|
| | I keep trying even if something is hard.[3] | .800 | 0.798 |
| | I complete my homework on time.[3] | .630 | 0.604 |
| **Engagement: cognitive engagement** | I try to learn more about the topics we cover in class.[3] | .741 | 0.733 |
| | I go through work that I do for math class and make sure it is right.[3] | .775 | 0.778 |
| | I think about different ways to solve a problem.[3] | .786 | 0.789 |
| | I try to connect what I'm learning to things I have learned before.[3] | .759 | 0.751 |
| | I try to understand my mistakes when I get something wrong.[3] | .708 | 0.74 |
| **Math enjoyment** | I often like to be challenged in math class.[3] | .642 | 0.661 |
| | I look forward to math class.[3] | .800 | 0.85 |
| | I enjoy learning new things in math class.[3] | .863 | 0.863 |
| | I want to understand what we are learning in class.[3] | .664 | 0.578 |
| | I feel good when I am in math class.[3] | .742 | 0.773 |
| | Math is interesting. | .644 | 0.741 |
| **Identity: identity as a "math person"** | I am just not good at mathematics.[4] | .509 | 0.543 |
| | I get good grades in mathematics.[4] | .694 | 0.641 |
| | I learn mathematics quickly.[4] | .853 | 0.863 |
| | In my mathematics class, I understand even the most difficult work.[4] | .784 | 0.777 |
| **Identity: others see me as a math person** | Parents/relative/friends[5] | .704 | 0.719 |
| | Mathematics teacher[5] | .572 | 0.571 |
| | Yourself[5] | .754 | 0.777 |
| **Math relevance** | Trying in mathematics is worth it because it will help me in the work that I want to do later on in life.[4] | .852 | 0.878 |
| | Learning mathematics is worthwhile for me because it will improve my career prospects.[4] | .863 | 0.882 |
| | Mathematics is an important subject for me because I need it for what I want to study later on.[4] | .824 | 0.827 |
| | I will learn many things in mathematics that will help me get a job.[4] | .794 | 0.827 |

*Note*. 1. Maloney & Matthews, 2020; 2. Hanson, 2017; 3. Fredricks et al., 2016; 4. OECD, 2013; 5. Cribbs et al., 2015.

### Exhibit 5. Model Fit of Measurement Invariance Models Over Time (Baseline–Outcome)

| Construct | Model | Chi-Square (df) | RMSEA | CFI | Diff Chi-Square (df) | Diff RMSEA | Diff CFI |
|---|---|---|---|---|---|---|---|
| **Belonging (3 factors)** | Configural | 778 (225) | .034 | .958 | | | |
| | Weak | 793 (234) | .033 | .968 | 15 (9) | .001 | -.010 |
| | Strong | 827 (243) | .033 | .966 | $34^{***}$ (9) | 0 | .002 |
| **Engagement (2 factors)** | Configural | 615 (192) | .032 | .974 | | | |
| | Weak | 629 (201) | .032 | .974 | 14 (9) | 0 | 0 |
| | Strong | 655 (210) | .032 | .973 | $26^{***}$ (9) | 0 | .001 |
| **Enjoyment** | Configural | 196 (47) | .039 | .982 | | | |
| | Weak | 219 (52) | .039 | .980 | $23^{***}$ (5) | 0 | .002 |
| | Strong | 338 (57) | .049 | .967 | $119^{***}$ (5) | -.010 | .013 |
| | Partial strong | 281 (56) | .044 | .973 | $62^{***}$ (4) | -.005 | .007 |
| **Identity (2 factors)** | Configural | 320 (64) | .044 | .964 | | | |
| | Weak | 325 (69) | .042 | .964 | 5 (5) | .002 | 0 |
| | Strong | 326 (74) | .040 | .965 | 1 (5) | .002 | .001 |
| **Relevance** | Configural | 95 (15) | .051 | .990 | | | |
| | Weak | 97 (18) | .046 | .990 | 2 (3) | .005 | 0 |
| | Strong | 100 (21) | .043 | .990 | 3 (3) | .003 | 0 |

*Note.* df = degrees of freedom; RMSEA = root mean square error of approximation; CFI = comparative fit; Diff = difference. $^{***}$ $p < .001$. All scales showed evidence of strong measurement invariance over time, except for the enjoyment scale that showed evidence of partial strong measurement invariance. The configural model is the baseline model with no constraints in the item parameters, the weak measurement invariance model tests the equality constraints in the item loadings, and the strong measurement invariance model tests the equality constraints in the item intercepts. The fit of these nested models was evaluated with the chi-square difference test. Because the chi-square difference test is highly sensitive to sample size, changes in CFI and RMSEA also were examined, with CFI changes of 0.01 or less and RMSEA changes of .015 considered acceptable (Chen, 2007; Cheung & Rensvold, 2002). The chi-square difference test as well as the changes in RMSEA and CFI for the strong measurement invariance model for the enjoyment construct indicate a decrease in model fit due to equality constraints in the intercepts. Further examination of the results indicated that the item “I want to understand what we are learning in class” shows different intercept values from baseline to outcome. A partial strong measurement invariance was fit to the data in which the intercepts for this item were allowed to differ, and the model showed good fit to the data. No modifications were made to this scale given that the item did not show differences in comparison between the treatment and control groups.

**Exhibit 6. Model Fit of Measurement Invariance Models Across Groups (Treatment–Control)**

| Construct | Model | Chi-Square (df) | RMSEA | CFI | Diff Chi-Square (df) | Diff RMSEA | Diff CFI |
|---|---|---|---|---|---|---|---|
| Belonging (3 factors) | Configural | 222 (102) | .057 | .973 | | | |
| | Weak | 237 (111) | .056 | .971 | 15 (9) | .001 | .002 |
| | Strong | 261 (120) | .057 | .968 | 24*** (9) | -.001 | .003 |
| Engagement (2 factors) | Configural | 216 (68) | .078 | .953 | | | |
| | Weak | 229 (76) | .075 | .952 | 13 (8) | .003 | .001 |
| | Strong | 238 (84) | .072 | .952 | 9 (8) | .003 | 0 |
| Enjoyment | Configural | 50 (18) | .071 | .983 | | | |
| | Weak | 51 (23) | .059 | .984 | 2 (5) | .012 | -.001 |
| | Strong | 61 (28) | .058 | .982 | 10 (5) | .001 | .002 |
| Identity | Configural | 97 (26) | .088 | .951 | | | |
| | Weak | 102 (31) | .081 | .951 | 5 (5) | .007 | 0 |
| | Strong | 106 (36) | .075 | .952 | 4 (5) | .006 | -.001 |
| Relevance | Configural | 18 (4) | .102 | .992 | | | |
| | Weak | 20 (7) | .073 | .993 | 2 (3) | .029 | .001 |
| | Strong | 21 (10) | .055 | .994 | 1 (3) | .018 | .001 |

*Note.* df = degrees of freedom; RMSEA = root mean square error of approximation; CFI = comparative fit; Diff = difference. *** $p < .001$. All scales showed evidence of strong measurement invariance over time. The configural model is the baseline model with no constraints in the item parameters, the weak measurement invariance model tests the equality constraints in the item loadings, and the strong measurement invariance model tests the equality constraints in the item intercepts. The fit of these nested models was evaluated with the chi-square difference test. Because the chi-square difference test is highly sensitive to sample size, changes in CFI and RMSEA also were examined, with CFI changes of 0.01 or less and RMSEA changes of .015 considered acceptable (Chen, 2007; Cheung & Rensvold, 2002).

# References

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://eric.ed.gov/?id=EJ772402

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. https://eric.ed.gov/?id=EJ651463

Cribbs, J. D., Hazari, Z., Sonnert, G., & Sadler, P. M. (2015). Establishing an explanatory model for mathematics identity. *Child Development, 86*(4), 1048–1062. https://eric.ed.gov/?id=EJ1069048

Fredricks, J. A., Wang, M. T., Schall, J., Hofkens, T. L., & Parr, A. (2016). Using qualitative methods to develop a survey measure of math and science engagement. *Learning and Instruction, 43*, 5–15. https://cadrek12.org/sites/default/files/Wangarticle2_0.pdf

Hanson, J. (2017). Determination and validation of the Project for Educational Research That Scales (PERTS) survey factor structure. *Journal of Educational Issues, 3*(1), 64–82. https://files.eric.ed.gov/fulltext/EJ1131721.pdf

Maloney, T., & Matthews, J. S. (2020). Teacher care and students' sense of connectedness in the urban mathematics classroom. *Journal for Research in Mathematics Education, 51*(4), 399–432. https://eric.ed.gov/?id=EJ1259551

OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* OECD Publishing. https://www.oecd.org/en/publications/2013/02/pisa-2012-assessment-and-analytical-framework_g1g27388.html

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231).

Widaman, K. F., & Olivera-Aguilar, M. (2023). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (2nd ed., pp. 367–384).





25533_07/24