The Algebra Concept Inventory: Creation and Validation with Students Across a Range of Math Courses in College

Claire Wladis
City University of New York

Kathleen Offenholley
City University of New York

Benjamin Sencindiver
University of Texas, San Antonio

Nils Myszkowski
Pace University

Geillan D. Aly
City University of New York

Even though algebraic conceptual understanding is recognized as a critical skill, existing larger-scale validated algebra assessments consist mostly of computational tasks, or only assess a very narrow range of conceptions in a smaller focused domain. Further, few instruments have been validated for use with college students. In this paper, we describe the creation and validation of an algebra concept inventory for college students. We describe how items were administered, revised, and tested for validity and reliability. Results suggest that algebraic conceptual understanding is a measurable construct, and that the instrument has reasonable validity and reliability. Revision and validation is ongoing; however, lessons learned thus far provide information about what conceptual understanding in algebra might look like and how it might be assessed.

Keywords: Algebra, conceptual understanding, concept inventory, assessment, validity

In college, needing to take algebra can be a barrier to degree completion (e.g., Adelman, 2006; Bailey et al., 2010), and extensive mathematics education research has documented K-12 students' difficulties with school algebra (e.g., Booth, 1988, 2011; Kieran, 1992). Struggles with basic algebra concepts learned in school also impact even those in higher-level college courses such as Calculus (e.g., Frank & Thompson, 2021; Stewart & Reeder, 2017). One reason students struggle with algebra is that algebra courses in college tend to focus on procedures disconnected from meaning-making (e.g., Goldrick-Rab, 2007; Hodara, 2011). While procedural fluency is important, it is critical to connect it with conceptual understanding (Kilpatrick, et al., 2001). Thus, there is a critical need to better understand and assess students' conceptions of algebra concepts. However, to date there are no widely-validated assessments that measure college students' conceptual understanding of algebra. Existing large-scale validated algebra assessments for K-12 tend towards computational skills, or focus on a narrow set of conceptions in a small subdomain. Tests of computational skills are often poor measures of understanding. Students may have robust conceptual understanding, but make smaller computational mistakes, especially if they have math or test anxiety (e.g., Ashcraft, 2002; Ashcraft & Kirk, 2001; Moran, 2016; Namkung et al., 2019). On the other hand, students may have little-or-no conceptual understanding, yet produce "correct" answers for "wrong" reasons (e.g., Aly, 2022; Erlwanger, 1973; Leatham & Winiecke, 2014).

We aim to address this gap by describing a first attempt to conceptualize, develop, and test college students' conceptual understanding in algebra using the *Algebra Concept Inventory* (*ACI*). This process continues, but we have chosen to write about results at this juncture with the hope they may be helpful for others interested in conceptualizing, measuring, and teaching conceptual understanding in algebra.

Literature Review

Several instruments have been created to test algebraic proficiency; however, none were designed to test a large body of algebraic concepts and conceptions. TIMMS and NAEP (Mullis, et al., 2020; National Center for Education Statistics, 2023) are widely validated at the international and national level, and contain some questions intended to assess conceptual understanding. There are also state-wide assessments that contain some questions intended to test conceptual understanding (e.g., Massachusetts Department of Elementary & Secondary Education, 2023; New York State Education Department, 2023). However, the main focus of all these instruments is computational skills.

There is one validated assessment that targets algebraic conceptual understanding in grades 1 to 5 (Ralston, et al., 2018), and one designed to assess a few specific algebraic concepts in middle school (Russell, 2019; Russell et al., 2009). Yet these instruments measure just a few conceptions, and were not designed for secondary or postsecondary students. As such, these often focus primarily on less complex or less abstract algebraic conceptions.

Some concept inventories have been developed that assess some student conceptions of algebraic concepts, but for students in more advanced courses only. For example, the Pre-calculus Concept Assessment (PCA) (Carlson, Oehrtman, & Engelke, 2010) and the Calculus Concept Readiness Instrument (CCRA) (Carlson, Madison, & West, 2010) explore some algebra concepts relevant to students in higher-level courses; while these have been tested through extensive cognitive interviews, larger-scale psychometric validation is still needed. Recently, researchers Hyland and O'Shea (2022) in Ireland generated a 31-item algebra concept inventory for college students, but it includes algebraic objects that would not be familiar to students in a first-year algebra course and has not yet been tested through cognitive interviews or psychometric analysis. Thus, an inventory that is valid for students starting in elementary algebra is needed, as well as more extensive large-scale psychometric testing of concept inventories more generally.

Method

A total of 402 unique items were developed and tested for the ACI. Items were administered to 18,234 students enrolled in all mathematics classes (except arithmetic) at a large urban community college campus. Data reported here were collected from spring 2019 to fall 2022, in eight separate waves. Data collection followed a common-item random groups equating design, which was selected because it allowed to investigate a large item pool while allowing a simultaneous calibration across multiple forms (de Ayala, 2009; Kolen & Brennan, 2004). For the first wave of testing, the last ten items on each form were anchor items, all taken from the National Assessment of Educational Progress (NAEP) grade 8 item bank. For subsequent waves, six anchor items were included: three of these were NAEP items and three were items that had performed well during the first wave of ACI testing. Each form had roughly 25 total items. Forms were randomly administered within in each class so there was no association between test form and class or instructor.

Just before answering inventory items, students were invited to participate in cognitive interviews, and paid for their time. In total, 135 interviews were conducted with students. Each was roughly 1-1.5 hours long and was structured as a "retrospective think-aloud" (Sudman et al., 1996). Research suggests that retrospective think-aloud protocols reveal comparable information to concurrent think-aloud protocols, and are less likely to have negative effects on task performance, particularly high-cognitive-load tasks (see e.g., van den Haak et al., 2003). Interviews were analyzed qualitatively to assess construct validity of the items, but there is insufficient space to report on that analysis here. Here we report only quantitative results.

We investigated each wave of the ACI through item-response theory analysis. First, items were dichotomized into pass-fail items using the response key. Then, two-parameter logistic models (Birnbaum, 1968) were estimated using marginal maximum likelihood (MML) on each wave, using the R package "mirt" (Chalmers, 2012). Because of planned missingness data collection design, the default number of model iterations was extended to allow for all models to converge successfully. Based on these models, we examined item parameters (difficulty and discrimination) and item information functions for item analysis, and computed person estimates using expected a posteriori (EAP) factor scores for convergent validity analysis. Reliability estimates were computed directly from IRT models. To investigate model fit, we computed item fit statistics, using the PV-Q1 statistic (and significance test) (Chalmers & Ng, 2017) for each item.

To investigate measurement invariance, we used multi-group IRT models and a model comparison approach. Because of the planned missingness design (and sometimes small observed subsample sizes), we used a piecewise DIF detection strategy (Thissen et al., 1993) that starts from a fully constrained model and drops constraints for each item separately. More specifically, with respect to each examinee characteristic considered, we first estimated a fully constrained model (where, across groups, item discriminations, difficulties, latent mean and variance are constrained to equality). Then, for each item, the same model was estimated, but with unconstrained item parameters (difficulty and discrimination), thus "temporarily" allowing differential item functioning (DIF) for the item. A likelihood ratio test was then performed to test if the model allowing DIF for the item had a better fit than the constrained model. This resulted in a series of tests of the significance of differential item functioning for all items. Because it is a multiple testing strategy, p-values were subsequently Bonferroni-corrected.

Validating the ACI

IRT Models: Item Discrimination and Difficulty

Results reported here were based on an item pool in which some items were dropped because they were deemed problematic (e.g., typographical errors; multiple correct answers); however, no items were dropped from analysis simply because of unsatisfactory IRT parameters. 2PL IRT models were run on all waves of data collection (Table 1). IRT models were run on all waves of data collection. While Rasch (or 1PL) models and 3PL models were also considered, 2PL models were chosen because unlike 1PL models, they allow discrimination to vary by item, and because they were considered more parsimonious, more useable for item selection (because item coefficients are more interpretable), and less prone to calibration errors than 3PL models due to their lower number of item parameters (San Martin et al., 2015).

Table 1. 2PL Model Coefficients Across all Eight Waves

Discrimination parameter	Proportion of Unique Items
>=0.65 "moderate"[a]	63.4%

>=1.35 “high”	31.3%
>=1.7 “very high”	18.5%
Difficulty parameter	Theta
mean	0.00
1st quartile	-0.85
median	-0.14
3rd quartile	0.63
Total number of unique items in 2PL models	399

[a] Characterizations of categories of discrimination parameters are taken from Baker (2001).

Discrimination is classified as “moderate” if it is ≥ 0.65, “high” if it is ≥ 1.35 and “very high” if it is ≥ 1.7 (Baker, 2001). Based on this, 63.4% of all items (253) have at least moderate, and roughly one-third have high or very high discrimination.

We also assessed item fit in the 2PL model for each wave using Chalmers’ $PV - Q_1$ test, because it performs better than other fit statistics at controlling Type I error (Chalmers & Ng, 2017) (Table 2).

Table 2. Measures of Item Misfit in 2PL IRT Models

	Number of Items With Significant[a] Misfit[b]	Total Number of Items	Percentage of Items With Significant Misfit
Wave 1	1	33	3.0%
Wave 2	5	125	4.0%
Wave 3	4	66	6.1%
Wave 4	3	72	4.2%
Wave 5	8	100	8.0%
Wave 6	5	99	5.1%
Wave 7	2	39	5.1%
Wave 8	0	31	0.0%
Total	28	565	5.0%

[a] Significant at the $\alpha = 0.05$ level
[b] Misfit as measured by Chalmers’ Chi-Square Statistic ($PV - Q_1$)

Only 5% of items were significantly misfitted by the 2PL models (for $\alpha = 0.05$), suggesting this is likely due to random variation.

Reliability

In IRT, the reliability of an item varies based on Theta, which represents the number of standard deviations above or below the mean an individual is on the measure of the latent trait. Table 3 shows various measure of reliability.

In Table 3 peak instrument values have excellent reliability ($R \geq 0.9$). There are also waves where excellent reliability ($R \geq 0.9$) can be obtained for values ranging from $\theta = [-2,7,2.2]$ (assuming a standard normal distribution of knowledge, this corresponds to satisfactory reliability for ~98% of examinees). Further, shorter tests can be constructed with only those items with the highest discrimination: for example, the 10 items with the best discrimination from Wave 1 yields a test with excellent reliability ($R \geq 0.9$) for $\theta = [-2,1]$.

Table 3. Reliability (R) for each wave of item administration of the ACI

	Theta at max info[a]	Info max[b]	R for info max[c]	theta w $R \geq$ 0.8	theta w $R \geq$ 0.9	Number of Items Tested
Wave 1	-1.4	26.4	0.96	[-2.8, 0.4]	[-2.4, -0.2]	33
Wave 2	-1.5	37.8	0.97	[-3.0, 2.1]	[-2.7, 0.9]	104
Wave 3	-0.6	24.3	0.96	[-2.3, 1.5]	[-1.8, 0.7]	57
Wave 4	-0.6	30.1	0.97	[-2.4, 2.1]	[-1.9, 1.2]	69
Wave 5	0.7	177.1	0.99	[-2.3, 2.9]	[-1.4, 1.8]	100
Wave 6	-0.6	105.3	0.99	[-1.7, 3.0]	[-1.0, 2.2]	99
Wave 7	-0.1	21.7	0.95	[-1.5, 1.8]	[-1.0, 1.1]	39
Wave 8	0.1	11.3	0.91	[-0.9, 1.2]	[-1.2, 0.3]	31

[a] info = 2PL IRT model information function

[bd] max = information function maximum for 2PL model

[e] $R = 1 - \frac{1}{Info}$

[c] expected reliability in Normal(0,1) ability distribution for 2PL models

Relationship Between ACI Score and Prior Algebra Course Completion: Convergent Validity

To explore convergent validity of the ACI, we explored the relationship between scores on the ACI (using theta scores from the 2PL model) to various measures of mathematics course level. For example, correlation of students' ACI scores with the level of algebra courses they have already successfully completed would be evidence of convergent validity. First, we consider linear regression models with level of student's course (where "level" is defined based on the algebra course pre-requisite requirements of the course) as the independent variable, and ACI score as the dependent variable (Table 4).

Table 4. Regression of course level (by algebra pre-requisite) in predicting theta scores from the 2PL model on the ACI, reference group: low

Course Level	Coefficient	SE	*p*-value (vs. low)	*p*-value (vs. high)
mid	0.347	0.014	0.000	0.000
high	0.700	0.017	0.000	

low = no algebra course prerequisite
mid = elementary algebra course prerequisite
high = intermediate algebra course prerequisite

In Table 4, differences in Theta score are significant ($p < 0.001$) for all pairwise comparisons. Students in "mid"-level courses scored on average 0.35 SD higher than those in "low"-level courses; and students in "high"-level courses scored on average 0.35 SD higher than those in "mid"-level courses (or 0.70 SD higher than in "low"-level courses). This provides strong evidence of convergent validity.

We also considered a more nuanced course sequence based on prerequisites (see Table 5).

Table 5. Sequence level of various courses in the sample, based on their prerequisites

Various elementary algebra courses	1
Various 100-level courses with an elementary algebra pre-requisite	2
Intermediate algebra courses	2
College algebra	2
Discrete math with intermediate algebra prerequisite	3
Precalculus	3
Math for elementary teachers with intermediate algebra prerequisite	3
Math for elementary teachers, second term	4
Advanced statistics with precalculus prerequisite	4
Introduction to geometry with precalculus prerequisite	4
Calculus I	4
Calculus II	5
Calculus III	6
Differential equations with Calculus II prerequisite	6
Linear algebra with Calculus II prerequisite	6
Abstract algebra	7

Rerunning linear regression models using this more refined set of levels again reveals a strong correlation between level and ACI score (Table 6).

Table 6. Regression of course position in longer mathematics curricular sequences (by classification given in Table 5) in predicting theta scores from the 2PL model on the ACI, reference group: sequence level 1

Course Position in Sequence	coeff	SE	*p*-value (vs. 1)	*p*-value (vs. 2)	*p*-value (vs. 3)	*p*-value (vs. 4)	*p*-value (vs. 5)	*p*-value (vs. 6)
2	0.504	0.017	0.000					
3	0.623	0.031	0.000	0.000				
4	0.888	0.023	0.000	0.000	0.000			
5	1.059	0.033	0.000	0.000	0.000	0.000		
6	1.232	0.041	0.000	0.000	0.000	0.000	0.000	
7	1.661	0.226	0.000	0.000	0.000	0.001	0.008	0.060

One of the largest gains (one half SD) was between sequence level 1 and 2 (see Table 6), which distinguishes between students who have or have not satisfied an elementary algebra prerequisite, providing further evidence of convergent validity, as the ACI is designed to focus on concepts relevant to elementary algebra specifically.

Differential Item Functioning: Measurement Invariance and Discriminant Validity

We also assessed potential differential item functioning (DIF) related to irrelevant examinee characteristics: race/ethnicity, gender, and English-language-learner status. This is an aspect of discriminant validity, as the ACI should measure algebraic conceptual understanding and not something else, like English literacy. Each wave was tested for DIF in three separate 2PL models: one for each characteristic. There was no consistent evidence of DIF on any of these factors. Only a negligible number of items had significant DIF for $\alpha = 0.05$ (using a Bonferroni correction for the number of tests within each model). Many items were tested in multiple waves, and none of these had significant DIF in more than one wave, suggesting that significant DIF in one wave for these items was likely due to random variation.

Limitations

The City University of New York, where this instrument was tested, is very diverse but not nationally-representative; however, this makes it useful for validation with marginalized students who have often been neglected in large-scale assessment validation. A current study is underway to validate the ACI on a larger national sample. The ACI has also only been validated with college students—further studies are necessary with younger students. The ACI has also been developed to make *diagnostic* judgements about *groups* of students—not high-stakes decisions for individuals—and thus we caution against that particular use of the ACI.

Discussion and Conclusion

Results from analysis suggest that algebraic conceptual understanding, as conceptualized by the items included on the ACI, is a measurable domain. IRT analysis indicated that a large proportion of items had good discrimination parameter estimates, suggesting that the final version of the ACI is likely to have an excellent ability to differentiate between students of various levels. Additionally, reliability was excellent for all waves, and results indicated that a shorter test could be constructed that would have excellent reliability for a large range of knowledge levels. The ACI also showed evidence of convergent validity, as students with higher algebra course prerequisites showed higher item success rates. Finally, only a negligible proportion of items showed differential item functioning with respect to race/ethnicity, gender, or English-language-learner status, indicating that the ACI had satisfactory measurement invariance with respect to these characteristics.

However, the ACI is only a first attempt at measuring algebraic conceptual understanding, and much more work needs to be done to map out in detail the various conceptions that students in different contexts hold of core algebra concepts, and determine how these can best be measured. The ACI provides only a single scale number; however, further work with cognitive diagnostic models on ACI items might provide more nuanced diagnostic information that could be particularly critical for instruction, by better modeling the complex layers of conceptions that students might have about various concepts in algebra. In reality, the kinds of knowledge that the ACI is trying to measure are quite complex, and capturing only a single score is, on its own, woefully inadequate if we hope to understand how students think algebraically and how various instruction and curriculum relate to this complex conceptual development. We see the ACI as just a first step in building out much more complex models of students' algebraic conceptions.

Acknowledgments

A grant from the National Science Foundation (#1760491) supported this research. Opinions reflect those of the authors and do not necessarily reflect those of the granting agency.

References

Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school through college* (ED490195). US Department of Education. http://www2.ed.gov/rschstat/research/pubs/toolboxrevisit/toolbox.pdf

Aly, G. (2022). Benny, Barbara, and the Ethics of EdTech. *Journal of Humanistic Mathematics*, *12*(2), 98–127.

Ashcraft, M. H. (2002). Math Anxiety: Personal, Educational, and Cognitive Consequences. *Current Directions in Psychological Science*, *11*(5), 181–185. https://doi.org/10.1111/1467-8721.00196

Ashcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. *Journal of Experimental Psychology: General*, *130*(2), 224–237. https://doi.org/10.1037/0096-3445.130.2.224

Bailey, T., Jeong, D. W., & Cho, S. (2010). Referral, Enrollment, and Completion in Developmental Education Sequences in Community Colleges. *Economics Of Education Review*, *29*(2), 255–270.

Baker, F. B. (2001). *The basics of item response theory*. ERIC.

Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability* (F. M. Lord & M. R. Novick, Eds.; pp. 395–479). Addison-Wesley.

Booth, L. R. (1988). Children's difficulties in beginning algebra. In *The ideas of algebra, K-12: 1988 Yearbook* (pp. 20–32). NCTM. http://elementaryalgebra.cmswiki.wikispaces.net/file/view/Childrens+Difficulties+in+Beginning+Algebra.pdf/142535729/Childrens+Difficulties+in+Beginning+Algebra.pdf

Carlson, M., Madison, B., & West, R. (2010). The calculus concept readiness (CCR) instrument: Assessing student readiness for calculus. *Eprint arXiv:1010.2719.*

Carlson, M., Oehrtman, M., & Engelke, N. (2010). The precalculus concept assessment: A tool for assessing students' reasoning abilities and understandings. *Cognition and Instruciton, 28*(2), 113-145. doi:https://doi.org/10.1080/07370001003676587

Chalmers, R. P. (2012). A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P., & Ng, V. (2017). Plausible-Value Imputation Statistics for Detecting Item Misfit. *Applied Psychological Measurement*, *41*(5), 372–387. https://doi.org/10.1177/0146621617692079

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. Guilford.

Erlwanger, S. H. (1973). Benny's conception of rules and answers in IPI mathematics. *Journal of Children's Mathematical Behavior*, *1*(2), 7–26.

Frank, K., & Thompson, P. W. (2021). School students' preparation for calculus in the United States. *ZDM – Mathematics Education*, *53*(3), 549–562. https://doi.org/10.1007/s11858-021-01231-8

Goldrick-Rab, S. (2007). What Higher Education Has to Say about the Transition to College. *Teachers College Record*, *109*(10), 2444–2481.

Hodara, M. (2011). *Reforming Mathematics Classroom Pedagogy: Evidence-Based Findings and Recommendations for the Developmental Math Classroom. CCRC Working Paper No. 27. Assessment of Evidence Series*. Community College Research Center, Columbia University; eric.

http://ezproxy.library.arizona.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED516147&site=ehost-live

Hyland, D., O'Shea, A. (2022). How Well Do High-Achieving Undergraduate Students Understand School Algebra?. *Can. J. Sci. Math. Techn. Educ.* 22, 818–834. https://doi.org/10.1007/s42330-022-00256-9

Kieran, C. (1992). *The learning and teaching of school algebra.*

Kilpatrick, J., Swafford, J., & Findell, B. (2001). The strands of mathematical proficiency. In *Adding it up: Helping children learn mathematics* (pp. 115–118). National Academies Press. http://books.google.com/books?hl=en&lr=&id=df7ZX4a8fzAC&oi=fnd&pg=PA1&dq=Adding+it+up:+Helping+children+learn+mathematics.&ots=jtZBgTu9cF&sig=WWHYCFi9heYlUsZP0oIvw1wJWXw

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*.

Leatham, K. R., & Winiecke, T. (2014). The case of the Case of Benny: Elucidating the influence of a landmark study in mathematics education. *The Journal of Mathematical Behavior*, *35*, 101–109. https://doi.org/10.1016/j.jmathb.2014.06.001

Massachusetts Department of Elementary & Secondary Education. (2017). Massachusetts comprehensive assessment system (MCAS). Retrieved from http://www.doe.mass.edu/mcas/

Moran, T. P. (2016). Anxiety and working memory capacity: A meta-analysis and narrative review. *Psychological Bulletin*, *142*(8), 831–864. https://doi.org/10.1037/bul0000051

Mullis, I. V., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). TIMSS 2019 international results in mathematics and science. *Retrieved from Boston College, TIMSS & PIRLS International Study Center website*: https://timssandpirls.bc.edu/timss2019/international-results.

Namkung, J. M., Peng, P., & Lin, X. (2019). The Relation Between Mathematics Anxiety and Mathematics Performance Among School-Aged Students: A Meta-Analysis. *Review of Educational Research*, *89*(3), 459–496. https://doi.org/10.3102/0034654319843494

National Center for Education Statistics. (2023). National assessment of educational progress (NAEP). Retrieved from https://www.nationsreportcard.gov/focus_on_naep/

New York State Education Department. (2017). Office of state assessment (OSA). Retrieved from http://www.nysedregents.org

Ralston, N. C., Li, M., & Taylor, C. (2018). The development and initial validation of an assessment of algebraic thinking for students in the elementary grades. *Educational Assessment*, *23*(3), 211–227.

Russell, M. (2019). Digital technologies: Supporting and advancing assessment practices in the classroom. In S. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and educational measurement* (pp. 224–242). Routledge.

Russell, M., O'Dwyer, L. M., & Miranda, H. (2009). Diagnosing students' misconceptions in algebra: Results from an experimental pilot study. *Behavior Research Methods*, *41*(2), 414–424.

San Martín, E., González, J., & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3PL model. *Psychometrika*, *80*, 450-467.

Stewart, S., & Reeder, S. (2017). Algebra Underperformances at College Level: What Are the Consequences? In S. Stewart (Ed.), *And the Rest is Just Algebra* (pp. 3–18). Springer International Publishing. https://doi.org/10.1007/978-3-319-45053-7_1

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In *Differential item functioning* (pp. 67–113). Routledge.

Van Den Haak, M., Jong, M. D., & Schellens, P. J. (2003). Retrospective vs. Concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behavior & Information Technology*, *22*(5), 339–351.