THE ALGEBRA CONCEPT INVENTORY FOR COLLEGE STUDENTS

Claire Wladis
City University of New York
cwladis@bmcc.cuny.edu

Kathleen Offenholley
City University of New York
koffenholley@bmcc.cuny.edu

Benjamin Sencindiver
University of Texas, San Antonio
benjamin.sencindiver@utsa.edu

Nils Myszkowski
Pace University
nmyszkowski@pace.edu

Geillan D. Aly
City University of New York
galy@bmcc.cuny.edu

There are currently no large-scale assessments to measure algebraic conceptual understanding, particularly among college students with no more than an elementary algebra, or Algebra I, background. Here we describe the creation and validation of the Algebra Concept Inventory (ACI), which was developed for use with college students enrolled in elementary algebra or above. We describe how items on the ACI were administered and tested for validity and reliability. Analysis suggests that the instrument has reasonable validity and reliability. These results could inform researchers and practitioners on what conceptual understanding in algebra might look like and how it might be assessed.

Keywords: Algebra and Algebraic Thinking; Equity, Inclusion, and Diversity; Undergraduate Education; Research Methods

Algebra can be a barrier to degree completion in college (e.g., Adelman, 2006; Bailey et al., 2010), and difficulties that K-12 students have experienced with algebra content has been extensively documented (e.g., Booth, 1988, 2011; Kieran, 1992). Understanding of key algebraic ideas has also been shown to impact college students in higher-level college courses like Calculus (e.g., Frank & Thompson, 2021; Stewart & Reeder, 2017). Algebra courses in college tend to focus on procedures disconnected from sense-making (e.g., Goldrick-Rab, 2007; Hodara, 2011), which may be one reason why college students in higher-level courses still struggle with algebraic ideas. It is important to connect procedural fluency with conceptual understanding (Kilpatrick, et al., 2001), and therefore, there is a critical need to better understand and assess students' algebra conceptions. However, there are not yet any widely-validated assessments to measure college students' algebraic conceptual understanding. Existing large-scale validated algebra assessments exist for K-12 students but focus primarily on computational skills, or only on a narrow subdomain of conceptions. Measures of computational skill are not necessarily valid measures of conceptual understanding, because 1) learners may have robust conceptual understanding, but make computational mistakes, particularly when they have math or test anxiety (e.g., Ashcraft, 2002; Ashcraft & Kirk, 2001; Moran, 2016; Namkung et al., 2019); or 2) learners may have little conceptual understanding, yet produce "correct" answers for the mathematically invalid reasons (e.g., Aly, 2022; Erlwanger, 1973; Leatham & Winiecke, 2014).

This paper describes how we have developed and tested college students' conceptual understanding in algebra using the *Algebra Concept Inventory* (*ACI*), in an attempt to address this gap. This process is ongoing.

Kosko, K. W., Caniglia, J., Courtney, S., Zolfaghari, M., & Morris, G. A., (2024). *Proceedings of the forty-sixth annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Kent State University.

Literature Review

While various algebraic proficiency instruments have been created, currently there are no widely-validated instruments that focus on a broad range of topics in algebraic conceptual understanding. TIMMS and NAEP (Mullis, et al., 2020; National Center for Education Statistics, 2023) have been widely validated nationally/internationally but have a broader focus and only contain a limited number of questions aimed at assessing algebraic conceptual understanding. There are also state-wide assessments that contain some items intended to measure conceptual understanding but that primarily focus on computational skills (e.g., Massachusetts Department of Elementary & Secondary Education, 2023; New York State Education Department, 2023). There are a few instruments that have been designed to measure a few specific algebra concepts in elementary or middle school (Ralston, et al., 2018; Russell, 2019; Russell et al., 2009), but these have not been tested with high school or college students, and the different population of interest means that the narrow range of conceptions do not include more complex or abstract conceptions that are critical to secondary and postsecondary mathematics.

Some concept inventories have been developed to assess algebraic conceptions relevant to calculus and other higher-level courses (Carlson, Oehrtman, & Engelke, 2010; Carlson, Madison, & West, 2010); however these instruments are not appropriate for students in lower-level courses such as elementary and intermediate algebra (or Algebra I/II in high school), and their focus is not on some of the core conceptions from these lower-level courses that may be particularly critical to algebraic reasoning. Further, while many of these have been tested extensively qualitatively, they have not to date published results of larger-scale psychometric validation. Recently, researchers Hyland and O'Shea (2022) in Ireland generated a 31-item algebra concept inventory for college students, but 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 algebra concept inventory that has been validated in large-scale data collection is sorely needed, particularly one that is appropriate for administration to students at all levels of prior algebra experience, and not just those in higher-level college courses.

Measuring Conceptual Understanding: Sample Item

There is insufficient space to describe the design of the ACI here, but it focuses on assessing specific conceptions of algebraic concepts (e.g., equivalence, syntactic meaning, algebraic properties, variable, function, covariation), rather than other skills like computation. For example, this item was designed to assess whether students can identify the existing syntactic structure of an algebraic expression vs. a procedure one might use to simplify the expression:

Sample item: Which of the following **best** describes the meaning of the expression $(2x + 3)(5x + 1)$ **as it is currently written**?

a. $2x$ is being multiplied separately by $5x$ and by 1, 3 is being multiplied separately by $5x$ and by 1, and these four results are being added together.
b. The result of adding $2x$ and 3 is being multiplied by the result of adding $5x$ and 1.

Method

A total of 402 unique items were developed and tested for the ACI. Items were administered to 7,658 students enrolled in all mathematics classes at the algebra level or above at a large urban

community college campus. Data were collected from spring 2019 to fall 2022, in eight waves. Data collection followed a common-item random groups equating design, selected because it allows investigation of a large item pool while allowing 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 were NAEP items and three were ACI items that had performed well during the first wave. Each form had roughly 25 items. Forms were randomly administered within each class to ensure 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. Each was roughly 1-1.5 hours long and structured as a "retrospective think-aloud" protocol (Sudman et al., 1996), which has been shown to reveal comparable information to concurrent think-aloud protocols, and is also less likely to have negative effects on task performance (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 that analysis here, where we focus on quantitative results.

To prepare data for item-response theory analysis, ACI items were dichotomized into correct/incorrect 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 the 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

Some items were dropped when issues were found during analysis (e.g., typographical errors; multiple correct answers); however, none were dropped due to unsatisfactory IRT parameters. 2PL IRT models were run on all waves (Table 1). We considered 1PL and 3PL models but chose

2PL models because they allow discrimination to vary by item and are more useable for item selection than 3PL models because item coefficients are more interpretable, and less prone to calibration errors 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 called as "moderate" if ≥ 0.65, "high" if ≥ 1.35 and "very high" if ≥ 1.7 (Baker, 2001). Based on these classifications, 63.4% of all items (253) have moderate or better, and roughly one-third have high or very high discrimination. Table 2 reports item fit for each wave using Chalmers' $PV - Q_1$ test, chosen because it performs better than other fit statistics at controlling Type I error (Chalmers & Ng, 2017). Only 5% of items were significantly misfitted by the 2PL models where $\alpha = 0.05$, which suggests that this is likely due to random variation.

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 ($PV - Q_1$)

Reliability

In IRT, Theta represents the number of standard deviations above or below the mean an

individual is on the measure of the latent trait, and the reliability of an item varies based on values of Theta. Peak information values for all waves (Table 3) have excellent reliability ($R \geq 0.9$). For various waves excellent reliability ($R \geq 0.9$) was 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). In addition, shorter tests can be constructed from a subset of 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

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 as predictor of ACI scores (2PL model)

Course Level[a]	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	

[a]reference group: low; low = no algebra course prerequisite; mid = elementary algebra course prerequisite; high = intermediate algebra course prerequisite; score is Theta score

Differences in scores in Table 4 are significant for all pairwise comparisons ($p < 0.001$). Scores for students in each level course were on average 0.35 SD higher than in the next lower course ("mid" vs. "low"; "high" vs. "mid"), providing 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

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

Table 6. Regression of more nuanced course level in predicting ACI score,

Course Position in Sequence	Coef.	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

The largest gain (one half SD) in Table 6 is between sequence level 1 and 2, or between students who have/have not satisfied an elementary algebra (Algebra I) prerequisite. This provides further evidence of convergent validity, because the ACI has been designed to focus on concepts relevant to elementary algebra specifically.

Differential Item Functioning: Measurement Invariance and Discriminant Validity

Differential item functioning (DIF) related to irrelevant examinee characteristics was also

analyzed, one subtype of discriminant validity (or whether the ACI measures algebraic conceptual understanding and not something else, like English literacy). Each wave was tested for DIF in three separate 2PL models: one each for race/ethnicity, gender, and English-language-learner status. There was no consistent evidence of DIF. 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 and not across models, which is overly conservative). 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 but not others for these items was likely due to random variation.

Limitations

The City University of New York (CUNY) where this instrument was tested is not nationally representative, and thus further research is needed to validate the ACI with less-diverse populations in other geographic areas; this research is currently underway with a larger national sample in the US. However, CUNY's diversity does make it a excellent candidate for initial validation with marginalized students who have often been neglected in large-scale assessment validation. Further studies are also necessary to determine whether the ACI may be valid for use with high school or middle school students. Finally, the ACI has been developed to make *diagnostic* judgements about *groups* of students—not high-stakes decisions for individuals—and thus the ACI should not be used alone to make high-stakes individual decisions such as course placement or successful course completion.

Discussion and Conclusion

This study suggests that algebraic conceptual understanding, as conceptualized by the items included on the ACI, is a measurable domain with reasonable validity and reliability. Item response theory (IRT) analysis resulting in large proportion of items with good discrimination parameter estimates, suggesting the ACI can differentiate well between students of various levels. Reliability was also excellent for all waves of data collection, and based on reliability estimates, even shorter tests can be constructed with excellent reliability for a range of levels of algebraic conceptual understanding. Students with higher algebra course prerequisites had higher ACI scores, providing evidence of convergent validity. Finally, differential item functioning analysis demonstrated that the ACI had satisfactory measurement invariance with respect to race/ethnicity, gender, or English-language-learner status.

However, the ACI in its current form is a summative measurement that provides only one measure of students' algebraic conceptual understanding. Future research could expand this to a more nuanced diagnostic tool that provides more detailed information about the specific conceptions that students have and what kinds of instructional approaches may be best adapted to students with different conceptions about various algebraic concepts. This work is ongoing, and includes in-depth qualitative analysis of student thinking to more comprehensively map out in more detail the various conceptions that students may hold of algebra concepts; work with cognitive diagnostic models on ACI items might provide more nuanced diagnostic information; exploration of different curricular materials and teaching techniques and the subsequent impact on the development of algebraic conceptual understanding. Our hope is that the ACI will also enable other practitioners and researchers to explore these questions as well.

Acknowledgments

This research was supported by a grant from the National Science Foundation (#1760491). 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=WWHYCFi9heYlUsZP0oIyw1wJWXw

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*. 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). 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.