Work in Progress: Validity and Reliability Testing of the Engineering Concept Assessment Modified for Eighth Grade

Dr. Kristin L. K. Koskey, University of Akron

Dr. Kristin Koskey is an Associate Professor in the LeBron James Family Foundation College of Education at The University of Akron. She holds a Ph.D. in Educational Research and Measurement and M.E. in Educational Psychology. Dr. Koskey teaches courses in evaluation, assessment, research design, and statistics. She also works as a psychometric consultant and serves on the Editorial Board for the journal of Psychological Assessment. Her work is published in leading journals such as Studies in Educational Evaluation, Journal of Applied Measurement, Journal of Mixed Methods Research, Journal of Experimental Education, International Journal of Qualitative Methods, and Educational and Psychological Measurement. Further, she has authored book chapters on Norming and Scaling for Automated Essay Scoring and Data-driven STEM Assessment. Dr. Koskey has secured grant funding from the Ohio Department of Education and National Science Foundation (NSF), as well as contributed to the evaluations on grants funded by the ODE, U.S. Department of State, and NSF.

Dr. Nicholas G. Garafolo, University of Akron

Dr. Nicholas G. Garafolo is a researcher in the broad area of thermo-fluids and aerospace, with an emphasis in advanced aerospace seals, near-hermetic fluid flows, and turbomachinery modal analysis. Dr. Garafolo currently holds a position as Assistant Professor at The University of Akron. Supporting the dissemination of his research activities, Dr. Garafolo has six journal manuscripts, over 30 conference papers and presentations, and $868,647 of total project funding. Prior to his appointment, Dr. Garafolo worked as a federal contractor, under the umbrella of a multi-million dollar contract, in space flight hardware research and development to NASA Glenn Research Center in Cleveland, Ohio. Dr. Garafolo was instrumental in developing a synergistic approach in the research and component modeling of elastomeric space seals for manned spaceflight; an asset to NASA and the development of advanced aerospace seals for the next generation of manned spacecraft. The unique problem necessitated a grasp of both fluid dynamics and material science, as well as experimental and computational analysis. As a DAGSI/Air Force Research Laboratory Ohio Student-Faculty Fellow, Dr. Garafolo gained experimental knowledge in structural dynamics of turbomachinery. In particular, his research on engine order excitation yielded insight into generating high cycle fatigue of turbomachinery using acoustic excitation.

Dr. Nidaa Makki, University of Akron

Dr. Nidaa Makki is an Associate Professor in the LeBron James Family Foundation College of Education at The University of Akron, in the department in Curricular and Instructional Studies. Her work focuses on STEM curriculum integration and science inquiry practices in middle and high school. She is a co-PI on an NSF funded project to investigate the impact of integrating engineering on middle school students' interest and engagement in STEM. She has also received funding to conduct teacher professional development in the areas of engineering education, problem based learning and inquiry instruction.

Dr. Wondimu Ahmed, University of Akron

Dr. Wondimu Ahmed is an Assistant Professor in the LeBron James Family Foundation College of Education at the University of Akron. He received his Ph.D. from University of Groningen, The Netherlands. His research focuses on motivation and emotions in education, particularly in STEM subjects.

Dr. Donald P. Visco Jr., University of Akron

Donald P. Visco, Jr. is the interim Dean in the College of Engineering at The University of Akron and Professor of Chemical & Biomolecular Engineering.

Mr. Uday Samreddy

©American Society for Engineering Education, 2017

Work-in-Progress: Validity and Reliability Testing of the Engineering Concept Assessment Modified for Eighth Grade

Providing experiences that promote interest in STEM during early adolescence is essential because evidence shows that middle school students' interest in science is an important predictor of later STEM career pursuit[1,2]. One way of providing such experiences is by integrating engineering into science through project-based learning activities. This research is part of a larger three-year study[3] testing the impact of an intervention integrating engineering into the forces and motion eighth-grade curriculum using a project-based approach. The intervention is aimed at increasing teachers' skills in applying inquiry-based teaching methods, as well as students' understanding of engineering design and interest in STEM. This work-in-progress paper reports on the process of the modification and reliability and validity testing of the Engineering Concept Assessment (ECA)[4] for eighth grade, here after referred to as the ECA-M8. The purpose is for the research team to obtain feedback on the modification process prior to implementing the measure to approximately 1800 students across 11 middle schools in during the third and final year of the larger study. The purpose of the ECA-M8 will be used as one indicator of intervention impact on student learning along with a performance assessment of understanding of engineering design, forces and motion concept assessment, and assessments of motivational outcomes including interest and self-efficacy in STEM. Another purpose of the ECA-M8 is for educators to use students' scores to inform instructional planning, as well as growth in understanding.

While there are established assessments for students' motivation in STEM[5,6] and students' understanding of forces and motion[7], the development of assessments of students' understanding of engineering design concepts is still in its infancy. Such assessments are needed because processes of design are considered to be the defining features of engineering[8]. One existing measure is the ECA developed by Daugherty et al.[4] targeted for high school students and teachers. A second measure is the Engineering Assessment developed by Harwell and colleagues[8] is targeted for grades fourth through eighth and consists of items from the previously established and validated TIMSS, NAEP, and AAAS. Harwell et al.[9] applied the Rasch model[10,11,12] to test the psychometric properties of the Engineering Assessment and found all 21 multiple-choice items (later reduced to 15 items to decrease length and time to complete the test) fit the Rasch model when completed by 168 sixth through eighth graders. Despite this finding, the assessment lacks items evaluating higher order thinking such as the ability to transfer engineering concepts to new design problems. Further, the final 15-item assessment produced only moderate reliability indices for the fourth and fifth grade student sample (.57) and sixth through eight-grade student sample (.71)[9].

Additional assessments of understanding of engineering design concepts are needed for different grade levels given that the curriculum and development levels differ across grades. In this work in progress paper, the process of the development of the ECA-M8 is overviewed and the preliminary item analysis results are reported for the multiple-choice portion of the assessment. The Rasch model[11,12] was adopted as the framework for evaluating the psychometric properties of the modified ECA. The Rasch model is commonly applied in constructing and evaluating attitudinal measures[13,14,15] and content tests in STEM[9] and provides information psychometric properties that Classical Test Theory approaches lack[16]. Further, the Rasch model is not sample dependent like Classical Test Theory approaches[16]. The following research questions were the main focus of this study with questions one through four addressed in this paper and data currently being collected and analyzed to address questions five and six.

1. What support is there for the content validity evidence for the ECA?
2. Did the items on the ECA fit the specifications of the Rasch model?
3. Did the ECA produce reliable item and person estimates?
4. Was there statistically significant differential item functioning (DIF) on the ECA across boys and girls?
5. What is the test-retest reliability of the ECA scores?
6. What is the reliability of the ECA rubric ratings when factoring in rater severity and item difficulty level as facets?

Assessment modification and validation process

Onwuegbuzie, Bustamante, and Nelson's[17] Instrument Development and Construct Validation (IDCV) process was adopted in the development and validation of the modified ECA. The Standards for Educational and Psychological Testing[18] recognize the importance of quantitative and qualitative evidence for informing the development and testing of assessments. The IDCV framework provides for both quantitative and qualitative evidences throughout the 10-phase mixed methods process. The 10-phase process adopted in this study is outlined in Appendix A.

Briefly, Phases 1 – 3 related to operationalizing the engineering design process using the state standards and Next Generation Science Standards[19] to guide item revision and development. The research team reviewed the existing ECA items to determine which items needed removal for misalignment with the standards being targeted or revision in wording for the eighth-grade level.

In Phase 4, three of the researchers conducted cognitive interviews using the modified assessment with five eighth-grade students during an after school program. Each interview lasted approximately 30 minutes. The students were asked to read each item aloud and talk aloud as they selected the answer or generated their answer to the essay items. The students were also asked if there were any words they did not know the meaning of or if there was anything confusing in the item stem or response options. The research team members corroborated their findings to inform further revisions.

Also in Phase 4 was the review of the assessment by a content expert in engineering, assessment expert, and expert in science education. The experts completed an online survey asking them to indicate whether each item aligned with the standard being targeted, was appropriate for the eighth-grade level, and the wording was clear or needed revision. Further, the experts were asked for feedback on the overall layout of the assessment. All experts rated the 13 multiple-choice and essay items as aligning to the standards. The two main revisions during this Phase involved changing a scenario in one multiple-choice item related to testing for skin cancer to be relevant to the students (the item was revised to focus on testing for strep throat) and more clearly deciphering among options provided for an item testing differentiating between practices of engineering and scientists.

The assessment expert and expert in science education also reviewed the rubric. The assessment expert suggested minor edits to the descriptors on the rubric and the science education expert found no revisions necessary. Additionally, the two researchers scored a sample of seven essay responses to the pre and post essay items to establish inter-rater reliability for the rubric ratings and identify further revisions needed through discussing disagreements in ratings.
Phases 5 -7 are currently being implemented involving field testing the ECA-M8 to a larger sample over two time points to evaluate the reliability and construct-related validity evidences. A dichotomous Rasch rating scale analysis was conducted for the pre-assessment multiple-choice responses and is reported in this paper as post-assessment data continues to be collected through February 2017.

The essay responses are currently being scored by undergraduate engineering students trained by the researchers using a cross-over method to provide for computation of inter-rater reliability and intra-rater reliability. The many-facet Rasch model[20,21] on the log-transformed data will be applied to the essay scores using FACETS[22] to produce parameter estimates for the multiple facets (students, items, raters) using maximum likelihood estimation. This model accounts for aspects of the measurement process that can systematically impact the estimation of a student's measure such as rater severity and item difficulty level.

During Phases 8 - 10, the instrument will be further revised based on the results of the quantitative psychometric analyses, as well as expert reviews. The ECA-M8 will then be administered to a larger sample of eighth grade students (~1800) to test for reliability and validity evidences for the revised instrument. The research team will engage in reflection on the development and validation process in Phase 10 to inform future research.

Description of the ECA-M8

The modified ECA consists of 13 multiple-choice items assessing basic understanding of engineering design concepts and one design problem testing their ability to transfer the concepts to a new design problem. Two design problem scenarios were developed, one for the pre-test and one for the post-test. Students were presented with five questions related to the design problem. Specifically, students identified the constraints of the problem, explained why or why not these interact, drew two designs that might be solutions, justified the selection of one to prototype, and described how to test the prototype.

Participants

A total of 561 eighth-grade students in an urban school district located in the mid-west completed the ECA-M8. This sample size exceeds the minimum of 30 persons for stable item calibrations and person measures within 1 logit applying a 95% confidence interval[23]. Eighty-nine percent of students in the district qualify for free/reduced priced lunches. Three hundred seventy four of the students were in the intervention group and 187 in the comparison group for the larger study. The students ranged from 12 to 15 years old with 95% of the sample reporting their age between 13 to 14 years old. Out of the 533 who reported their gender, 243 were boys and 290 were girls. Of the 519 reporting their ethnicity, the majority of the sample was made up of 44% (n = 233) were Black followed by 33% (n = 175) White, 10% (n = 56) Asian or Pacific Islander.

Preliminary multiple-choice item analysis results

Descriptive statistics for the total scores are reported in Table 1. Noteworthy is that the total scores were expected to be low (36% correct) since the intervention was not yet implemented.

Table 1

Descriptive Statistics for Total Scores by Gender and Research Group (N = 561)

Descriptive	Gender		Group		Overall Sample
Statistic	Boys	Girls	Intervention	Comparison	
n	243	290	374	187	561
M	4.66	4.88	5.07	4.13	4.76
SD	2.37	2.38	2.33	2.34	2.37
Mdn	5.00	5.00	5.00	4.00	5.00
Skewness	0.13	0.24	0.23	0.13	0.18
Kurtosis	– 0.26	– 0.17	– 0.13	– 0.17	– 0.21

An Independent Samples *t*-test showed that the intervention group scored statistically significantly higher than the comparison group at baseline, $t_{559} = 4.50$, $p < .001$. The mean difference was 0.94 points, yielding a 0.40 effect size. According to the What Works Clearinghouse standards published by IES[24], effect sizes > 0.25 indicate that the groups are non-equivalent and that statistical adjustment is insufficient for addressing the baseline differences. As a result, a sub-sample will be drawn from the comparison group to create an equivalent comparison group at baseline when comparing in future analyses on this outcome. To provide for a continuum of person abilities when conducting the Rasch analyses for instrument development purposes, the two groups were collapsed for the dichotomous Rasch rating scale analyses. Significant differences between boys and girls in STEM through high school continue to be documented[25]. No statistically significant difference was found in total scores by gender in their ECA-M8 scores, $t_{531} = –1.03$, $p = .303$.

The dichotomous Rasch rating scale analysis was applied and is expressed as:

$$\mathrm{P}[X_{ni} = 1] = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}$$

where the probability that person n will score correctly (1) on item i is a function of the constant exponent equal to 2.71828 (natural log) raised to the difference between a person's ability (β_n) and the item difficulty (δ_i).

The item analysis indices are reported in Appendix B including the infit and outfit mean square (MNSQ) and standardized (Z_{STD}) indices indicating the fit of the data to the Rasch model. The

MNSQ is the transformed residuals representing the difference between the observed and predicted with an expected MNSQ value of 1[16].

Item difficulty level ranged from –1.18 to 0.96, indicating additional items at a higher difficulty level such as the essay design problem items are needed, as one of the purposes is to assess student growth from lower to higher-level understanding. Item separation was 4.94, yielding sufficient separation of more than four[1] distinct groups of items along the measure. The item separation reliability of 0.96 indicated the ordering of items along the continuum would likely replicate for similar samples[16]. Person separation was 1.04 with a separation reliability of 0.52, likely an artifact of the homogeneous sample of students not yet exposed to the intervention. The person mean of – 0.72 (*SD* = 1.07) logits was below the set item mean of zero.

All of the items had positive point-biserial measures ranging from 0.24 to 48. Bond and Fox[16] recommend fit values between 0.30 and 1.30 as acceptable with values greater than 1.30 misfitting and below 0.30 overfitting the model. Based on this criterion, all 13 items are within the acceptable parameters (see Appendix B). However, as recommended by Smith, Schumacker, and Bush[26] and Smith[27], item misfit was also identified using Z standardized statistics (Z_{STD}). Any item yielding a $Z_{STD} \geq \pm 2.33$ ($\alpha = 0.01$) was identified as not fitting the Rasch model. As reported in Appendix B, item 7 and item 13 had statistically significant Z_{STD} infit values, $p <$ 0.01 and require further inspection of the response patterns. In examining the content of the items, item 7 related to deciphering between practices of engineers and scientists as to whether one or the other or both "use models to test designed systems and recognize strengths and limitations." About 13% selected scientific practice, 42% selected engineering practice, and 37% selected practice of both science and engineering. Item 13 required the students to determine which of three machine designs was the most difficult to adapt given a change from sorting ping pong balls to golf balls. About 30 % selected Sink or Swim, 23.3% selected Chute Sorter, and 36.3% selected Conveyor Belt.

Rasch-Welch *t*-test was applied to test for statistically significant DIF by gender using a logistic regression model expressed as:

$$\text{Log}\ [\text{pnij}/\text{Pni(j-1)}] = \beta n - \delta gi - \text{F}j$$

βn is the ability level of student n, δgi is the difficulty of student i in classification g, and Fj is equal to 0 since all items are in the same item-grouping with only two groups[28]. The DIF results are reported in Appendix B. Item 11 yielded a DIF size of –1.16 logits, above the criterion of .50 [27]; however, this DIF was not statistically significant. No items yielded statistically significant DIF by gender.

Next steps

Further analysis of the response patterns for items 7 and 13 will be conducted. Also, a partial credit Rasch rating scale analysis will be conducted adding the essay item scores on the pre and post test data to determine whether (a) these two items continue to misfit, (b) the person

[1] H = (4G +1)/3 whereby G is the standard deviation over the average measurement error[29].

reliability increases, and (c) if the item ordering (i.e., based on item difficulty level) is stable across administrations. Based on these findings and further content and assessment expert review, additional revisions will be completed prior to administering the assessment to a larger sample during the final year of the project. Finally, as recommended by the reviewers, the essay items will be counterbalanced in the final year 3 of the project such that half of the sample is randomly assigned to complete Form A at the pre-test and the other half Form B.

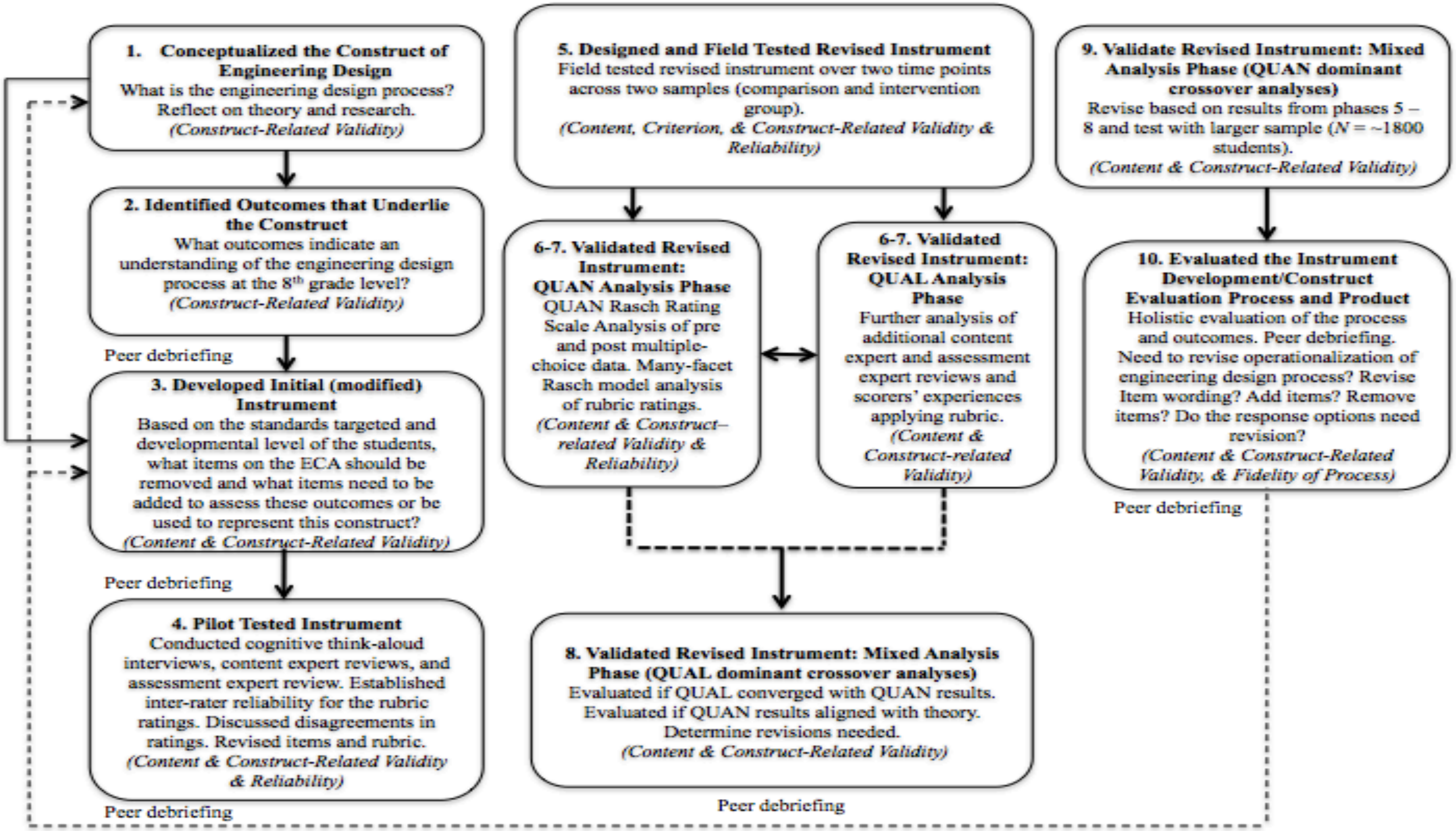
Bibliography

1. Sadler, P. M., Sonnert, G., Hzari, Z., & Tai, R. (2012). Stability and volatility of STEM career interest in high school: A gender study. *Science Education, 96*, 411-427. doi:10.1002/sce.21007
2. Tai, R. H., Qi-Liu, C., Maltese, A. V., & Fan, X. (2006). Planning early for careers in science. *Science, 312* (5777), 1143-1144. Retrieved from https://doi.org/10.1126/science.1128690
3. Makki, N., Garafolo, N. G., Halasa, K., Ahmed, W., Koskey, K. L. K., & Visco, D. (in press). Exploring the engineering design process through computer aided design and 3D printing. Manuscript accepted for publication in *Science Scope*.
4. Daugherty, J., Custer, R. L., Brockway, D., & Spake, D. A. (2012). Engineering Concept Assessment: Design and development (AC 2012-2987). American Society for Engineering Education.
5. Greene, B. A. (2015). Measuring cognitive engagement with self-report scales: Reflections from over 20 years of research. *Educational Psychologist, 50*, 14-30. doi:10.1080/00461520.2014.989230
6. Unfried, A., Faber, M., Stanhope, D. S., & Wiebe, E. (2015). The development and validation of a measure of Student Attitudes Toward Science, Technology, Engineering, and Math (S-STEM). *Journal of Psychoeducational Assessment,* 1-18.
7. American Association for the Advancement of Science (2017). *Science assessment.* Washington, DC.
8. Moore, T. J., Glancy, A. W., Tank, K. M., Kersten, J. A., Smith, K. A., & Stohlmann, M. S. (2014). A framework for quality K-12 engineering education: Research and development. *Journal of pre-college engineering education research*, *4*(1), 2.
9. Harwell, M., Moreno, M., Phillips, A., Selcen Guzey, S., Moore, T. J., Roehrig, G. H. (2015). A study of STEM assessments in engineering, science, and mathematics for elementary and middle school students. *School Science and Mathematics, 115*, 66-74.
10. Andrich, D. (1978). A rating formulation for ordering response categories. *Psychometrika, 43*, 357-374.
11. Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Dankmarks Paedagogiske Institut.
12. Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (expanded ed.). Chicago: University of Chicago Press.
13. Boone, W. J., Townsend, J. S., & Starver J. (2011). Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. *Science Education, 95*, 258-280. doi:10.1002/sce.20413

14. Eggert, S., & Bögenholz, S. (2009). Students' use of decision-making strategies with regard to socioscientific issues: An application of the Rasch partial credit model. *Science Education, 94*, 230-258. doi:10.1002/sce.20358
15. Liu, X. (2010). *Using and developing measurement instruments in science education: A Rasch modeling approach.* Charlotte, NC: Information Age Publishing.
16. Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
17. Onwuegbuzie, A. J., Bustamante, R. M., & Nelson, J. A. (2010). Mixed research as a tool for developing quantitative instruments. *Journal of Mixed Methods Research, 4*(1), 56-78. doi:10.1177/1558689809355805
18. American Educational Research Association, American Psychological Association, National Council on Measurement in Education [Joint Committee on Standards for Educational and Psychological Testing] (2014). *Standards for educational and psychological testing.* Washington, DC. AERA.
19. National Research Council (2013). *Next Generation Science Standards: For states, by states.* Washington, DC: The National Academies Press.
20. Linacre, J. M. (1994). *Many-facet Rasch measurement.* Chicago, IL: MESA Press.
21. Linacre, J. M., & Wright, B. D. (2002). Understanding Rasch measurement: Construction of measures from many-facet data. *Journal of Applied Measurement, 4*, 486-512.
22. Linacre, J. M. (2006). *FACETS* [Computer software program]. Chicago, IL: MESA press.
23. Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.
24. Institute of Education Sciences (2014). *What Works Clearinghouse: Procedures and standards handbook (version 3.0).* Retrieved from https://ies.ed.gov/ncee/wwc/Handbooks
25. Institute of Education Sciences (2015, February). *Gender differences in science, technology, engineering, and mathematics (STEM) interest, credits earned, and NAEP performance in the 12th grade.* Retrieved from https://nces.ed.gov/pubs2015/2015075.pdf
26. Smith, R. M., Schumacker, R. E., and Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*, 66-78.
27. Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement, 1*, 199-218.
28. Linacre, J. M. (2016). *A user's guide to WINSTEPS MINISTEP Rasch-model computer programs* [Program manual 3.92.0]. Retrieved from winsteps.com
29. Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago, IL: MESA Press.

Appendix A

10-Phase Process of Instrument Modification and Validation



1. Conceptualized the Construct of Engineering Design
What is the engineering design process? Reflect on theory and research.
(Construct-Related Validity)

2. Identified Outcomes that Underlie the Construct
What outcomes indicate an understanding of the engineering design process at the 8th grade level?
(Construct-Related Validity)

3. Developed Initial (modified) Instrument
Based on the standards targeted and developmental level of the students, what items on the ECA should be removed and what items need to be added to assess these outcomes or be used to represent this construct?
(Content & Construct-Related Validity)

4. Pilot Tested Instrument
Conducted cognitive think-aloud interviews, content expert reviews, and assessment expert review. Established inter-rater reliability for the rubric ratings. Discussed disagreements in ratings. Revised items and rubric.
(Content & Construct-Related Validity & Reliability)

5. Designed and Field Tested Revised Instrument
Field tested revised instrument over two time points across two samples (comparison and intervention group).
(Content, Criterion, & Construct-Related Validity & Reliability)

6-7. Validated Revised Instrument: QUAN Analysis Phase
QUAN Rasch Rating Scale Analysis of pre and post multiple-choice data. Many-facet Rasch model analysis of rubric ratings.
(Content & Construct–related Validity & Reliability)

6-7. Validated Revised Instrument: QUAL Analysis Phase
Further analysis of additional content expert and assessment expert reviews and scorers' experiences applying rubric.
(Content & Construct-related Validity)

8. Validated Revised Instrument: Mixed Analysis Phase (QUAL dominant crossover analyses)
Evaluated if QUAL converged with QUAN results. Evaluated if QUAN results aligned with theory. Determine revisions needed.
(Content & Construct-Related Validity)

9. Validate Revised Instrument: Mixed Analysis Phase (QUAN dominant crossover analyses)
Revise based on results from phases 5 – 8 and test with larger sample (N = ~1800 students).
(Content & Construct-Related Validity)

10. Evaluated the Instrument Development/Construct Evaluation Process and Product
Holistic evaluation of the process and outcomes. Peer debriefing. Need to revise operationalization of engineering design process? Revise Item wording? Add items? Remove items? Do the response options need revision?
(Content & Construct-Related Validity, & Fidelity of Process)

Onwuegbuzie et al.'s[17] IDCV process adapted for the development and testing of the Engineering Concept Assessment Modified for eighth grade (ECA-M8). QUAN = quantitative. QUAL = qualitative. Phases 1 – 5 completed. Currently engaging in Phases 6 - 8. Peer debriefing occurred at key decision phases such as to inform the item and structural validity and verify revisions to items. Content experts in middle school science education and engineering, an assessment expert, and the project evaluator served as peer debriefers depending on the phase of revisions.

Appendix B

Rasch Rating Scale Analysis and Differential Item Function Analysis Results

Item #	Next Generation Standard Aligned to Item[1]	Rasch Measure (in logits)	*SE*	Z_{STD} Fit Statistic[1]		MNSQ Fit Statistic		*pb*	*p*-value[3]	DIF Results	
				Infit	Outfit	Infit	Outfit			DIF	*t*-value[4]
6	ETS1.B	0.96	.11	1.13	1.22	1.07	1.12	0.24	.20	0.96	.02
3	ETS1.A	0.46	.10	0.87	1.46	1.04	1.10	0.30	.28	0.47	.03
2	ETS1.A	0.38	.10	–0.42	–0.22	0.98	0.98	0.36	.29	0.39	.04
4	ETS1.A	0.18	.10	0.03	0.55	1.00	1.03	0.36	.33	0.18	.04
10	ETS1.C	0.18	.10	0.74	0.42	1.03	1.02	0.34	.33	0.18	.04
1	ETS1.A	0.16	.10	1.03	1.63	1.04	1.09	0.33	.33	0.16	.04
12	ETS1.B, ETS1.C	0.07	.10	–2.00	–1.93	0.93	0.90	0.43	.35	0.07	.05
13	ETS1.B, ETS1.C	–0.04	.10	2.45[2]	1.88	1.09	1.09	0.30	.37	–0.03	.05
9	Science and Eng. Practices	–0.13	.09	–0.17	–0.60	0.99	0.97	0.39	.39	–0.13	.06
7	Science and Eng. Practices	–0.32	.09	–3.51[2]	–3.31	0.89	0.86	0.48	.43	–0.31	.06
5	ETS1.A	–0.35	.09	–0.65	–0.71	0.98	0.97	0.41	.43	–0.35	.07
8	Science and Eng. Practices	–0.37	.09	–1.15	–1.41	0.96	0.94	0.43	.43	–0.36	.07
11	ETS1.A, ETS1.B	–1.18	.10	0.08	–0.49	1.00	0.97	0.43	.60	–1.16	.15

Note. Items ordered from most difficult to least difficult. [1]ETS1.A = Defining and delimiting engineering problems. ETS1.B = Developing possible solutions. ETS.C: Optimizing the design solution. [2]Item misfits, $Z_{STD} \geq 2.33$, $p < .01$. [3]Percentage correct. Rounded to nearest whole. [4]No items statistically significantly DIF by gender.