Assessing the Reliability of a Chemical Engineering Problem-solving Rubric when Using Multiple Raters

Mr. Timothy Ryan Duckett, Acumen Research and Evaluation, LLC

T. Ryan Duckett is a research associate with Acumen Research and Evaluation, LLC., a program evaluation and grant writing company that specializes in STEM and early childhood education. He is a PhD student in the Research and Measurement department at the University of Toledo.

Prof. Matthew W. Liberatore, University of Toledo

Matthew W. Liberatore is a Professor of Chemical Engineering at the University of Toledo. He earned a B.S. degree from the University of Illinois at Chicago and M.S. and Ph.D. degrees from the University of Illinois at Urbana-Champaign, all in chemical engineering. His current research involves the rheology of complex fluids as well as active learning, reverse engineering online videos, and interactive textbooks. His website is: http://www.utoledo.edu/engineering/chemical-engineering/liberatore/

Uchenna Asogwa, University of Toledo

Uchenna Asogwa is a graduate student of Chemical Engineering at the University of Toledo. He earned a B.S. degree from the University of Benin, Nigeria in chemical engineering. His current research involves the reverse engineering online videos as well as rheology of complex fluids.

Dr. Gale A. Mentzer, Acumen Research and Evaluation

Gale A. Mentzer, PhD, the owner and director of Acumen Research and Evaluation, has been a professional program evaluator since 1998. She holds a PhD in Educational Research and Measurement from The University of Toledo and a Master of Arts in English Literature and Language—a unique combination of specializations that melds quantitative and qualitative methodologies. She and has extensive experience in the evaluation of projects focused on STEM education including evaluations of several multi-million dollar federally funded projects. Previously she taught graduate level courses for the College of Education at The University of Toledo in Statistics, Testing and Grading, Research Design, and Program Evaluation.

Dr. Amanda Portis Malefyt, Trine University

Amanda Malefyt is currently Chair and Associate professor in the McKetta Department of Chemical and Bioprocess Engineering at Trine University. She received her bachelor's degree from Trine (formerly Tri-State) University and Ph.D. from Michigan State University. Her research interests include engineering education and nucleic acid therapeutics.

Assessing the reliability of a chemical engineering problem-solving rubric when using multiple raters

Abstract

This evidence-based practice paper discusses the preliminary validation of a project modified version of the Promoting Problem Solving Proficiency in First Year Engineering (PROCESS). The full rating plan required four raters to use the PROCESS to assess the problem-solving ability of ~70 engineering students randomly selected from two undergraduate cohorts at two Midwest universities. The many-facet Rasch measurement model has the psychometric properties to determine if there are any characteristics other than problem-solving that influence the scores assigned to students, such as rater bias or differential item functioning. Prior to implementing the full rating plan, the analysis examined how raters interacted with the six items on the modified PROCESS when scoring a random selection of 20 students' solutions to one textbook homework problem. Follow up inter-rater reliability meetings enabled rater discussion of rationale for discrepancies observed in the ratings. Differences in conceptions of the latent construct of problem-solving were resolved by recourse to the theoretical framework that informed the development of the PROCESS. This iterative process resulted in substantial increases in construct validity and measurement reliability when raters completed another round of assessment. Evidence indicated that raters increased their understanding of how rating scale categories related to levels of the latent construct. This paper describes the impacts and benefits this method of psychometric evaluation of rater-mediated assessments hold for the implementation of the full rating plan of student outcomes, as well as for the field of engineering education more broadly.

Introduction

Engineers require precision and reliability in the tools they use to conduct research. For instance, the optimal design of planning vessels that transport goods around the world relies on the consistency of repeated particle image velocimetry measurements of flow characteristics around a ship [1, 2]. Yet much work is still required to develop tools for use in engineering education that meet the same rigorous standards of accuracy and repeatability when it comes to the assessment of student outcomes [3-5].

The attempts in engineering education to meet the demands of accountability and to provide assurances in the assessment of student knowledge have been marked by several components. There are institutes and committees comprised of engineering professors from across the country who develop and validate cognitive/declarative knowledge exams that serve as summative course assessments [6]. In response to the call for more robust learning outcomes, many science and engineering departments have integrated professional development programs that bolster faculty familiarity with course evaluation concepts [7, 8]. Incorporating multiple types of student assessment in classroom instructional design has been found to increase proficient practice in the field [9].

Methods of student assessment often incorporate rater-mediated assessment [10-13]. These methods of assessing student knowledge move beyond traditional notions of student grades that

are just the calculation of correct responses divided by total possible items on a formative test. In rater-mediated assessment, student performance on a given task (e.g. presentation, homework solution, concept paper) is judged by a rater along any number of domains through the use of a rating scale [14]. The inclusion of these additional components provides the prospect of more nuanced and detailed student assessments, but also the threat of greater inconsistency. Efforts need to be made to ensure that the rubric used for rating students represents the intended construct. This task necessitates the development of a continuum where students can be placed according to their possession of less or more of the latent construct attempted to be measured indirectly through their performance on the given task. Rating scales need to distinguish between distinct levels of performance. Raters need to be consistent in their use of the rating scales. Figure 1 provides a model of rater-mediated assessment of problem-solving. This model shows that students are placed along the continuum of problem-solving ability by raters using rating scale judgments of student performance on a set of tasks.

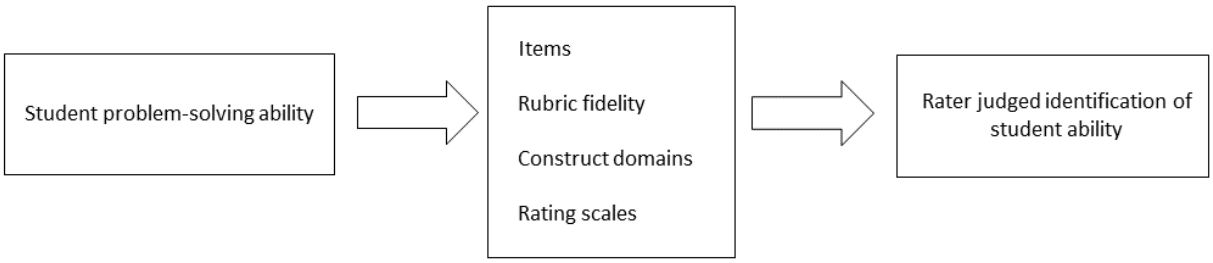

Figure 1. A model for rater-mediated assessment. Adapted from [14].

The qualitative levels defined by each category on a rating scale represent unequal intervals along the latent construct [15]. The conveniently adopted ordinal level ratings given to the qualitative categories (e.g. a "score" of 1 for "Inadequate," "2" for "Acceptable," etc.) need to be converted into linear measures before they can be used in any meaningful way as proxy measures of student ability levels. For example, one would be hard pressed to argue that "Disagree" minus "Strongly Disagree" equals an integer value of 1. Therefore, each facet of the assessment situation as shown in Figure 1, above, becomes a parameter that is estimated in a many-faceted measurement approach. Figure 2 describes the many-facet Rasch model (MFRM) developed by Linacre based upon the Rasch measurement paradigm [15-17]. This approach treats the assigned ordinal ratings in an assessment as the outcomes of the linear combination of the parameters. A comparison of the empirical variance encountered during parameter estimation with the level of measurement error expected by the model indicates how well the data fit the model. Unlike other item response and classical test theory traditions that try to fit measurement models to the data, the Rasch model is built on the fundamental measurement property of invariance: the measurement of persons must be independent of particular items used for measuring (item-invariant person measurement) and the calibration of items must be independent of particular persons used for calibration (person-invariant item calibration) [18, 19]. The task of measurement in the Rasch paradigm becomes an investigation of how well a particular data set adhere to the principles of invariant measurement embedded in an ideal-type model [14, 20].

$$\log\left(\frac{Pnijjk}{Pnijk\text{-}1}\right) = Bn - Di - Cj - Fk$$

where:

P_{nijk} is the probability of examinee *n*, when rated on item *i* by judge *j*, being awarded a rating of *k*.

P_{nijk-1} is the probability of examinee *n*, when rated on item *i* by judge *j*, being awarded a rating of *k-1*.

B_n is the ability of examinee *n*.

D_i is the difficulty of item *i*.

C_j is the severity of judge *j*.

F_k is the extra difficulty overcome in being observed at the level of category *k*, relative to category *k-1*.

Figure 2. Equation for the many-facet Rasch model.

The purpose of this paper is to estimate the reliability of rater-mediated assessments of undergraduate engineering student problem-solving. Latent constructs such as problem-solving ability and content mastery comprise the domain of learning outcomes and variables of interest in the field of education. The MFRM was developed using philosophical principles similar to those that underpin the physical measures in engineering. Use of the MFRM can determine the fairness and objectivity of the estimations of student problem-solving by accounting for all of the aspects of the measurement process that can introduce error into that estimation, such as poorly functioning items, ill-defined rating categories, and differing levels of rater severity. Establishing reliability in rater-mediated assessments provides evidence that the scores obtained on the test actually represent the latent construct instead of being an artifact of rater discrepancies [21]. This paper argues that the MFRM provides necessary evidence toward the validity of inferences that can be made regarding student learning outcomes in engineering education.

Methods

Participants

A total of 113 students were enrolled in an undergraduate Material and Energy Balance chemical engineering course as part of a control cohort (23 students; 22% female) and a treatment cohort (93 students; 41% female) at two Midwest Universities. Table 1 shows different distributions for highest mathematics courses completed by cohort. This discrepancy can be explained as a consequence of the course sequence occurring in the sophomore year for the control cohort (fall and spring semesters) compared to the spring semester of the freshmen year for the treatment cohort.

Table 1

Highest completed mathematics course by cohort type

	Control		Treatment	
	Count	%	Count	%
Calculus 1	2	9%	61	68%
Calculus 2	12	52%	11	12%
Calculus 3	7	30%	11	12%
Differential Equation	2	9%	3	3%
> Differential Equation	0	0%	4	5%

Instrument

The PROCESS was used to score students' homework problem solutions [10]. The PROCESS was developed using several theoretical frameworks that consider the conceptual, analytical, and phenomenological process demands and cognitive skills involved in problem solving [22]. PROCESS was modified to assess the problem-solving process for solved handwritten homework problems, which differs from its original use where participants' solutions were collected on tablets and custom software to see erasing and other details [23, 24]. The tool was modified to suit material and energy balance problems. The modified PROCESS consists of a 6-stage rubric assessing: Problem definition, Representing the problem, Organizing information, Calculations, Solution completion and Solution accuracy. Each item in the revised PROCESS consists of four scaling levels ranging from 0 to 3 with the following categories to rate student performance on each of the six the stages of problem solving: missing, inadequate, acceptable, and accurate. Any identification regarding group identity was removed prior to scoring and replaced with a project-assigned ID number to maintain privacy and to mask group membership from raters.

A complete rating plan was proposed where four raters would use the PROCESS tool to score all solutions submitted by all students from both cohorts. The four raters consisted of one chemical engineering faculty member, one high school science teacher, and one graduate and one undergraduate student in chemical engineering. All students completed ten traditional textbook problems during the respective courses.

Analyses

Initial inter-rater reliability was assessed in line with best practices as a means to evaluate how consistently raters measured student problem-solving ability [25]. The first assessment involved the PROCESS scores that five raters assigned to 20 randomly selected students for one textbook problem. An additional chemical engineer faculty member joined the four raters above to provide a benchmark reference point. Figure 3 presents a portion of the problem that was purposefully selected for piloting the use of the rubric. This specific problem was chosen in part because the research team decided it was of average difficulty and representative of the ten textbook problems assigned.

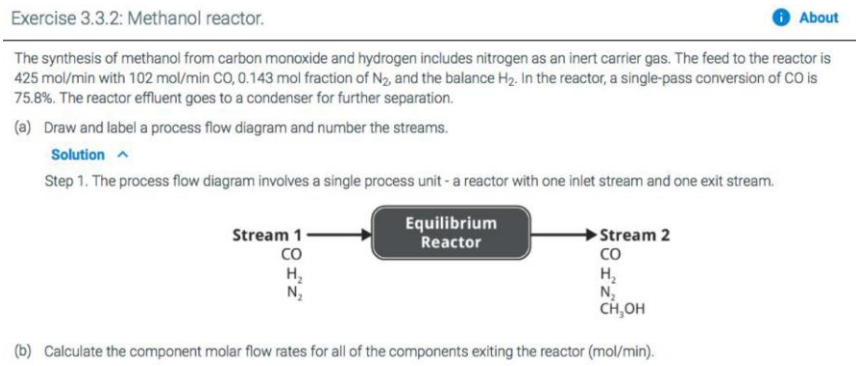

Figure 3. Example of the textbook problem used to rate student problem-solving ability

The FACETS [16] computer program was used to produce parameter estimates for the facets involved in the rater-mediated assessment (rating scale function, rater severity, item difficulty, etc.). Subsequently, qualitative focus group meetings were conducted where raters deliberated rationale for their ratings and their understanding of the underlying continuum of problem-solving. Discrepancies in ratings were resolved by recourse to the theoretical framework of the problem-solving cycle that informed the development of the PROCESS [22, 26]. Raters were then rescore that problem in light of their refined understanding of the latent trait and function of PROCESS. Those results were then analyzed in the same manner as before using the FACETS program to estimate parameters. The resulting logit scores were rescaled to conform to the original scale of 0 points (a rating of "missing" for all six PROCESS items, representing the lowest problem-solving ability level) through 18 points (a rating of "accurate" for all six PROCESS items, representing the highest problem-solving ability level).

These interval level measures were then used to calculate Cohen's kappa and intraclass correlation coefficients as extra measures of inter-rater reliability in addition to the standard errors and fit statistics provided by FACETS. Several types of descriptive statistics were calculated to assess the inter-rater reliability of the four raters using the PROCESS. The goal was to ensure that all of the raters used the rating scale consistently so that differences in student performance can be attributed to different problem-solving abilities and not a result of receiving a rating from a more or less severe rater.

Results

The problem-solving continuum developed by FACETS as a result of the parameter estimations of rater severity, student ability level, item difficulty, and rating scale function are displayed in Figure 4. Students with higher scores indicate more advanced problem-solving skills, such that student 4835 was identified as exhibiting the most advanced problem-solving skills and student 1874 as the least advanced. For the PROCESS Item column, the higher the item is on the "ruler," the more difficult it is for students to answer it correctly. Therefore, "Final Solution Accuracy" was the most difficult item, with "Representing the Problem" and "Final Solution Completion" being the easiest items (i.e. students scored the highest on these items). The Rater column places the more severe raters (i.e. gave the lowest scores—Raters 3 and 5) higher on the ruler and the more lenient raters (i.e. awarded the highest scores—Rater 4) lower on the ruler.

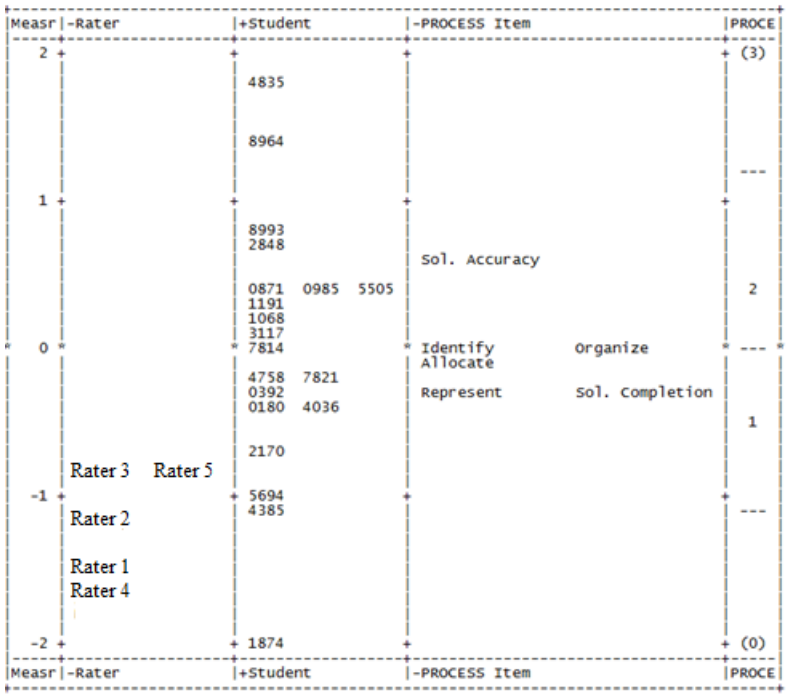

Figure 4. Yardstick representation of student ability level, produced in FACETS

Figure 5 provides more descriptive statistics regarding the estimation of the rater severity parameter. The raters are ordered in rows from most to least severe, with their overall measure being reported in the fifth column. Of particular interest is the Strata statistic of 2.48 (highlighted in yellow). This indicates that there were two distinct groups of raters, a more severe group and a more lenient group.

Several types of rater agreement could be achieved [16]. If the desire is to have raters agree exactly with each other then we would expect the third to last column in Figure 5 (Exact Agree. Obs. %) to be greater than 90%. This would mean that raters were agreeing on exact scores for student performance on individual PROCESS items greater than 90% of the time. As can be seen, in this instance this is not the case. Of greater concern to most measurement contexts is the determination of similar leniency/severity in rater assessments. This is reported by the 'Reliability (not inter-rater)' statistic, highlighted in red in Figure 5, and calculated by taking [1 – Separation Reliability], [1 – 0.72 = 0.28], with numbers closer to 0 being best. While there are no hard and fast guidelines, 0.28, in conjunction with other evidence, suggests that the raters were rating with different levels of severity. Similarly, the null hypothesis for the "Fixed (all same) chi-square tests" shown in the third row from the bottom of Figure 5, assumes that all raters are rating the same. The significant statistic chi-square value (highlighted in green, Figure 5) indicates that we must reject this null hypothesis, providing further evidence that raters are rating with different levels of severity and therefore the raters are bringing different interpretations of the rating scale into their scoring of student problems

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Exact Agree. Obs %	Exact Agree. Exp %	N	Rater
282	120	2.4	2.51	-1.17	.13	1.51	2.9	1.26	1.3	1.05	.34	65.8	52.6	5	Rater 5
287	120	2.4	2.55	-1.25	.13	.96	-.2	1.24	1.1	.83	.31	61.3	53.5	3	Rater 3
303	120	2.5	2.66	-1.54	.14	.92	-.4	1.05	.3	.90	.32	65.0	56.0	2	Rater 2
311	120	2.6	2.71	-1.71	.15	.82	-.9	.75	-1.0	1.09	.40	69.8	56.9	1	Rater 1
313	120	2.6	2.73	-1.76	.15	.78	-1.2	.66	-1.4	1.10	.41	70.2	57.2	4	Rater 4
299.2	120.0	2.5	2.63	-1.48	.14	1.00	.0	.99	.1		.36			Mean (Count: 5)	
12.6	.0	.1	.09	.24	.01	.26	1.5	.25	1.1		.04			S.D. (Population)	
14.0	.0	.1	.10	.27	.01	.29	1.7	.28	1.3		.05			S.D. (Sample)	

Model, Populn: RMSE .14 Adj (True) S.D. .19 Separation 1.37 Strata 2.16 Reliability (not inter-rater) .65
Model, Sample: RMSE .14 Adj (True) S.D. .23 Separation 1.61 Strata 2.48 Reliability (not inter-rater) .72
Model, Fixed (all same) chi-square: 14.9 d.f.: 4 significance (probability): .00
Model, Random (normal) chi-square: 3.2 d.f.: 3 significance (probability): .37
Inter-Rater agreement opportunities: 1200 Exact agreements: 797 = 66.4% Expected: 663.0 = 55.2%

Figure 5. Descriptive statistics for parameter estimation of rater severity.

Figure 6 models the discrepancy in rater severity by plotting the range of precision in the estimation of rater scores (calculated as Rater Measure ± (2 x S.E.)). This shows that Rater 5 when most lenient was still significantly more severe than Rater 4 at their most severe. This indicates that assessment of student problem-solving in this initial example was influenced by measurement error introduced as a result of different levels of rater severity or, rather, different interpretations by raters of what constituted each level of the rubric. Thus, a student would get a different problem-solving ability score dependent on which rater assessed their assignment.

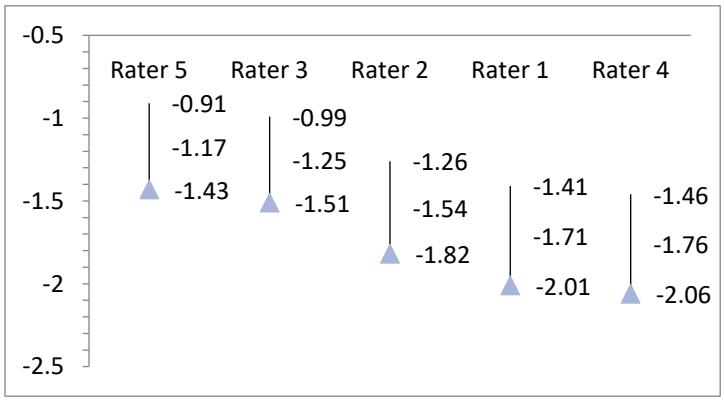

Figure 6. Calculation of the range of rater severity from FACET parameter estimation

Further diagnosis revealed some of the overarching areas of disagreement. For example, Table 2 reveals statistically significant bias regarding how Rater 5 scored the first PROCESS item, "Identify the Problem," and Rater 3's rating of the second item, "Represent the problem." The scores the raters gave on those respective items across all 20 students they rated were statistically significantly lower than expected by the model given the estimated student problem-solving ability level and item difficulty. The t-statistics were computed to test the hypothesis that there was no bias present in the ratings besides what was to be expected by measurement error. This analysis indicated bias in how these raters scored these particular items; specifically, they were harsher when scoring these items compared to other PROCESS items.

Table 2
Bias in rater interaction with PROCESS items

Rater	PROCESS Item	Observed Score	Expected Score	Bias Size (log odds units)	Model standard error	*t-statistic*	d.f.	p-value
5	Identify the problem	36	47.2	-0.82	0.25	-3.29	19	0.0038
3	Represent the problem	44	50.8	-0.66	0.28	-2.34	19	0.0302

Comparisons between how pairs of raters scored specific items can lead to fruitful conversations regarding the characteristics of responses that belong to each rating category on the items, i.e. "What does an answer to Item 1, Identify the problem, need to look like to be considered in the acceptable category (score of 2)?" The data in Table 3 suggest that the raters employed different understandings of item 1, "Identify the problem." This was the greatest source of disagreement leading to different levels of rater severity. For example, row 1 reports that Rater 1 (overall severity estimate of -0.48 logits on item 1) has a statistically significantly more lenient understanding/rating of item 1, Identify the problem when compared to Rater 5's (0.84 logit) more strict position on that item, t (35) = -2.55, p =.015. This indicates that a student solution that was assessed by Rater 5 was likely to receive a significantly lower score on the "Identify the problem" item than if that same solution was assessed by Rater 1.

Table 3
Pairwise comparison of rater discrepancies in scores assigned to PROCESS items

PROCESS Item	Rater Pair	Contrast	S. E.	t-statistic	d. f.	p-value
Identify the problem	Rater 1 - Rater 5	-1.32	0.52	-2.55	35	0.015
Identify the problem	Rater 2 - Rater 3	-1.54	0.64	-2.39	33	0.023
Identify the problem	Rater 2 - Rater 5	-2.01	0.63	-3.19	32	0.003
Identify the problem	Rater 3 - Rater 4	1.73	0.76	2.27	32	0.030
Identify the problem	Rater 4 - Rater 5	-2.21	0.75	-2.94	30	0.006

The qualitative meeting between raters examined the discrepancies highlighted above, in addition to those found in the raw scores raters provided to some of the students, displayed in Table 4. Highlighted cells show areas of considerable discrepancy in ratings that could potentially represent different understandings of the underlying construct and its measurement. It can be seen that the majority of the ratings provided by the raters were similar across the PROCESS items for most of the students. The solution provided by student 4036, a moderate performing student, provided difficulties that led to fruitful conversations about the different characteristics of responses in the "inadequate" and "acceptable" rating scale categories. The meeting offered the opportunity for the raters to clarify any fundamental disagreements or misunderstandings pertaining to the latent construct of problem-solving ability.

Table 4

Discrepancies in use of rating scale categories

Rater	Student ID	Identify Problem score	Represent the Problem score	Organize Knowledge Score	Allocate Resources score	Final solution Completion score	Final Solution Accuracy score
Rater 1	5694	1	3	1	0	0	0
Rater 2	5694	1	2	1	0	0	0
Rater 3	5694	3	2	1	0	1	0
Rater 4	5694	1	2	1	0	0	0
Rater 1	8993	3	3	3	3	3	3
Rater 2	8993	3	2	2	2	3	2
Rater 3	8993	3	3	3	3	3	3
Rater 4	8993	3	3	3	3	3	3
Rater 1	3117	3	3	3	2	3	1
Rater 2	3117	3	2	3	3	3	1
Rater 3	3117	3	3	3	3	3	3
Rater 4	3117	3	3	3	3	2	1
Rater 1	4036	3	3	3	2	3	1
Rater 2	4036	3	2	1	1	2	1
Rater 3	4036	2	3	3	3	2	3
Rater 4	4036	3	3	1	1	1	1

Following the qualitative meeting, the four primary raters were asked to rescore the student solutions for the "methanol reactor" problem described above. Inter-rater reliability statistics were computed to assess the extent to which the raters understood and scored "problem-solving ability" in a consistent manner with each other. The results in Table 5 report the two forms of Cohen's kappa that were calculated. The second column reports the standard Cohen's kappa for absolute agreement. This statistic quantifies how many instances of exact agreement occurred for the ratings (e.g. both raters would have to give a particular student the same rating on a specific item). This method is best suited to determine absolute level of agreement, essentially treating the ratings as binary outcomes, and therefore only has limited applicability here. It is included in the present study just to provide a baseline for comparison. Column 3 reports the quadratic weighted kappa statistics. These take into account the nature of ordered categories and adjust for the fact that adjacent categories are more alike than non-adjacent (i.e. ratings of 0 from one judge and 1 from another are more similar than ratings of a 0 and a 3). Moderate levels of agreement (kappa statistic in the range of .60 - .79) mark half of the rater relationships; specifically all of those relationships that do not involve Rater 3. Table 5 highlights the need to follow up with Rater 3 to discuss their understanding of the latent variable and potentially provide additional training on their understanding of the rating scale categories that comprise the assessment tool.

Table 5

Cohen's kappa coefficient estimates based on recalibrated ratings

Pair	κ value (absolute)	κ value (quadratic weights)
Rater 1 - Rater 2	0.384	0.707
Rater 1 - Rater 3	0.115	0.573
Rater 1 - Rater 4	0.465	0.743
Rater 2 - Rater 3	0.100	0.421
Rater 2 - Rater 4	0.395	0.792
Rater 3 - Rater 4	0.254	0.536

Additionally, intraclass correlation coefficients (ICC) were computed to assess how similar the group of raters as a whole (rather than rater pairs) scored problem-solving ability for the student solutions. This correlation shows the reproducibility of the measurement of student problem-solving ability. Table 6 reports the results of the intraclass correlation coefficient computed using the interval level estimates produced by FACETS as the measure. The ICC coefficient was calculated using the initial assessment of the raters and then again using the ratings of the second round of assessment. Specifically, a two-way mixed effects, multiple-raters model was employed. The initial ICC coefficient of .844, indicates good reliability. Yet, when taking into consideration the 95% CI [.570, .973], there was a wide range from very low moderate to excellent reliability in the raters' scores. The ratings from the post-qualitative meeting show appreciable increases in reliability and a narrower confidence interval. Provided the raters score student solutions with a similar conceptualization of problem-solving ability, then adding more student solutions will help reduce the high standard error caused by the relative small sample size in this analysis.

Table 6
Intraclass correlation coefficient for rater agreement

Ratings	Intraclass Correlation	95% Confidence Lower Bound	95% Confidence Upper Bound	F value	df1	df2	Sig
Pre-qualitative meeting	0.85	0.57	0.97	22.57	5	15	< .001
Post-qualitative meeting	0.90	0.69	0.98	34.99	5	15	< .001

Figure 7 maps the rater-mediated student scores used to calculate the post-qualitative meeting ICC coefficient. Data came from six randomly selected students. This plot provides a snapshot of how raters differed in the sum scores given to each of the students. It should be noted that Rater 3 shows significant separation from the other three raters on three of the six assessments (A, C, and F), while Rater 2 was drastically lower for the ratings on student B. This provides further explanation for the variation in the 95% CI of the ICC coefficient.

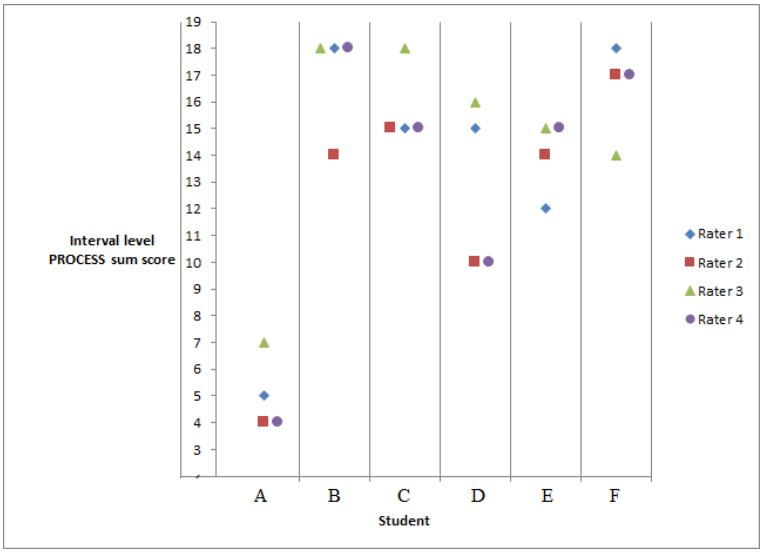

Figure 7. Plot of the student problem-solving ability level used for the ICC coefficient

Discussion

This study estimated the reliability of scores from a rubric designed to measure chemical engineering problem-solving ability. The analyses mark an important step in the validation of the PROCESS itself which has only been validated previously using traditional correlational techniques. The many-facet Rasch model (MFRM) was used to explore a set of rater-mediated data. This evaluative approach and choice of measurement models was designed to meet the increasing demands of accountability in engineering, and in this case specifically chemical engineering, education [3, 6, 27]. The models in the Rasch measurement paradigm are particularly well suited to the task of evaluating the defensibility of measures pertaining to the assessment of student learning outcomes. The demands of specific objectivity, as set out by Rasch, require person-free item calibration and item-free person measures [18]. The Rasch models also expect the measures resulting from data to meet other requirements similar to those maintained for the physical measures that define the field of chemical engineering. Examples include the requirements of monotonicity and local independence. These conditions demand that items with increasing levels of difficulty require increasing presence of the latent variable in order for an individual to succeed on that item/receive a higher rating [28].

The initial findings show promise for the validity of the measures of problem-solving ability produced by the PROCESS. Qualitative meetings discussed the raters' conceptions of the latent construct and how the rating scales mapped progress along the continuum of lower to greater levels of the construct. The conversations produced a more stable understanding of the thresholds of each of the rating categories, e.g. the hallmarks of an "inadequate" (rating of 1) response to a PROCESS item and at what point that response became "acceptable" (rating of 2). Future analyses will monitor the function of the rating scale categories as different chemical engineering problems are scored. The inclusion of more rater-mediated assessments will make for more precise parameter estimations and therefore lead to student assessments that more accurately represent actual student ability. Fortunately, the MFRM evaluation process is iterative in nature and can (and should) be conducted as assessments are ongoing [25]. This evaluation process can identify sources of measurement error in any of the facets estimated, including the parameter of rater severity. Discrepancies in use of the rating scale on the PROCESS tool can provide opportunities for additional training. These evaluative steps can increase not only the accuracy of the scores among the raters, but also the precision of those scores in measuring the latent construct, provided the raters maintain fidelity in their use of the rating scale rubric and the operationalization of the problem-solving framework.

Reports suggest that there will only be an increase in the call for authentic, meaningful measures of student outcomes in engineering programs as the 21st Century proceeds [29]. Novel methods of engaging students in the content and methods of engineering appear promising [11, 30, 31]. The validity of the pedagogical interventions and the inferences that can be drawn from resulting measures will be enhanced through the use of robust measurement and evaluation techniques. Engineering educators who demand measures as sturdy as the measures used to build the machines that cultivate alternative energy [32] and fuel the next modes of transportation [33] need to implement a rigorous system of evaluation of their pedagogical assessments through the use of a measurement model that makes such demands on the data. To that end, the implementation of Rasch measurement models will provide robust validation for the measures of

student learning outcomes, which in turn can improve course curricula by accurately targeting domains and transferable skillsets critical to the development of this generation's chemical engineers.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. DUE 1712186. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work was completed within the framework of University of Toledo IRB protocol 202214.

Bibliography

[1] J. Y. Lee, B. G. Paik, and S. J. Lee, "PIV measurements of hull wake behind a container ship model with varying loading condition," *Ocean Engineering,* Article vol. 36, no. 5, pp. 377-385, 2009.

[2] L. Wan-zhen, G. Chun-yu, W. Tie-cheng, X. Pei, and S. Yu-min, "Experimental study on the wake fields of a ship attached with model ice based on stereo particle image velocimetry," *Ocean Engineering,* Article vol. 164, pp. 661-671, 2018.

[3] E. Crawley, J. Malmqvist, S. Ostlund, and D. Brodeur, in *Rethinking engineering education: The CDIO Approach*, 2007.

[4] J. E. Froyd, P. C. Wankat, and K. A. Smith, "Five major shifts in 100 years of engineering education," *Proceedings of the IEEE,* vol. 100, no. Special Centennial Issue, pp. 1344-1360, 2012.

[5] M. Towns, "Guide to developing high-quality, reliable, and valid multiple-choice assessments," *Journal of Chemical Education,* vol. 91, no. 9, pp. 1426-1431, 2014.

[6] T. Holme, "Assessment and Quality Control in Chemistry Education," *Journal of Chemical Education,* Article vol. 80, no. 6, p. 594, 2003.

[7] M. Emenike, J. R. Raker, and T. Holme, "Validating chemistry faculty members' self-reported familiarity with assessment terminology," *Journal of Chemical Education,* vol. 90, no. 9, pp. 1130-1136, 2013.

[8] X. Wang and S. Hurley, "Assessment as a Scholarly Activity?: Faculty Perceptions of and Willingness to Engage in Student Learning Assessment," *Journal of General Education,* vol. 61, no. 1, pp. 1-15, 01/01/ 2012.

[9] M. H. Towns, "Developing Learning Objectives and Assessment Plans at a Variety of Institutions: Examples and Case Studies," *Journal of Chemical Education,* Article vol. 87, no. 1, pp. 91-96, 01// 2010.

[10] G. Sarah Jane, D. Jennifer Van, B. Lisa, and M. Beshoy, "Process Analysis as a Feedback Tool for Development of Engineering Problem Solving Skills," Atlanta, Georgia, 2013/06/23, Available: https://peer.asee.org/22372

[11] M. F. Brian, S. Natalie, and A. K. James, "How We Know They're Learning: Comparing Approaches to Longitudinal Assessment of Transferable Learning Outcomes," New Orleans, Louisiana, 2016/06/26, Available: https://peer.asee.org/25493

[12] K. L. K. Koskey, G. G. Nicholas, M. Nidaa, A. Wondimu, P. V. Donald, Jr., and S. Uday, "Work in Progress: Validity and Reliability Testing of the Engineering Concept Assessment Modified for Eighth Grade," Columbus, Ohio, 2017/06/24, Available: https://peer.asee.org/29189
[13] M. G. Chad, L. Quinn, F. F. Brian, and H. Liv, "Determining Reliability of Scores from an Energy Literacy Rubric," Seattle, Washington, 2015/06/14, Available: https://peer.asee.org/23820
[14] G. Engelhard Jr and S. Wind, *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge, 2017.
[15] J. M. Linacre and B. D. Wright, "Understand Rasch Measurement: Construction of Measures from Many-facet Data," *Journal of Applied Measurement*, vol. 3, no. 4, pp. 486-512, 01/01/ 2002.
[16] J. Linacre, *Many-facet Rasch measurement*. Chicago, IL: MESA Press, 1989.
[17] D. Andrich, "Controversy and the Rasch model: a characteristic of incompatible paradigms?," *Medical Care*, vol. 42, no. 1, pp. I7-I16, 2004.
[18] G. Rasch, "On Specific Objectivity: An attempt at Formalizing the Request for Generality and Validity of Scientific Statements," *Danish Yearbook of Philosophy*, vol. 14, pp. 58-94, 1977.
[19] B. Wright, "Sample-free Test Calibration and Person Measurement. ETS Invitational Conference on Testing Problems," MESA Research Memorandum1967.
[20] B. D. Wright and M. Stone, *Measurement essentials*, 2 ed. Wilmington, Delaware: Wide Range, Inc., 1999, p. 221.
[21] E. W. Wolfe and J. E. Smith, "Instrument development tools and activities for measure validation using Rasch models: part I-instrument development tools," *Journal of applied measurement*, vol. 8, no. 1, pp. 97-123, 2007.
[22] S. J. Grigg and L. C. Benson, "A coding scheme for analysing problem-solving processes of first-year engineering students," *European Journal of Engineering Education*, vol. 39, no. 6, pp. 617-635, 2014.
[23] S. J. Grigg, J. Van Dyken, L. Benson, and B. Morkos, "Process analysis as a feedback tool for development of engineering problem solving skills," in *ASEE Annual Meeting*, Atlanta, 2013, p. 6505.
[24] S. J. Grigg and L. Benson, "Promoting problem solving proficiency in first year engineering process assessment," in *ASEE Annual Meeting*, Seattle, WA, 2015.
[25] J. Linacre, "Judging plans and facets," *Mesa Research Note*, vol. 3, 1997.
[26] J. E. Pretz, A. J. Naples, and R. J. Sternberg, "Recognizing, defining, and representing problems," in *The psychology of problem solving*, vol. 30, 2003, pp. 3-30.
[27] K. Vilonen and J. Linnekoski, "LEARNING OUTCOMES AND ASSESSMENT AS COURSE DEVELOPMENT TOOLS," in *IGIP-SEFI Annual Conference in Trnava*, 2010.
[28] R. Perline, B. D. Wright, and H. Wainer, "The Rasch model as additive conjoint measurement," *Applied Psychological Measurement*, vol. 3, no. 2, pp. 237-255, 1979.
[29] P. H. Meckl, M. H. Williams, C. Percifield, M. E. Cardella, M. T. Harris, and L. H. Jamieson, "Taking Stock: Progress toward Educating the Next Generation of Engineers," in *American Society for Engineering Education*, 2012: American Society for Engineering Education.

[30] M. W. Liberatore, D. W. Marr, A. M. Herring, and J. D. Way, "Student-created homework problems based on YouTube videos," *Chemical Engineering Education,* vol. 47, no. 2, pp. 122-132, 2013.
[31] M. W. Liberatore, C. R. Vestal, and A. M. Herring, "YouTube Fridays: Student led development of engineering estimate problems," *Advances in Engineering Education,* vol. 3, no. 1, pp. 1-16, 2012.
[32] C. Niya, Y. Rongrong, C. Yao, and X. Hailian, "Hierarchical method for wind turbine prognosis using SCADA data," *IET Renewable Power Generation,* Article vol. 11, no. 4, pp. 403-410, 2017.
[33] S. Wu, Y. Du, and S. Sun, "Transition metal dichalcogenide based nanomaterials for rechargeable batteries," *Chemical Engineering Journal,* Article vol. 307, pp. 189-207, 2017.