Statistical Analysis and Report on Scale Validation Results for the Engineering Ethical Reasoning Instrument (EERI)

Peter Wesley Odom, Purdue University - Department of Engineering Education

Wesley is a PhD student in Engineering Education at Purdue University. His primary research interests surround assessment technologies, the psychology of student learning of STEM subjects, ethics, and international community development.

Dr. Carla B. Zoltowski, Purdue University-Main Campus, West Lafayette (College of Engineering)

Carla B. Zoltowski is an assistant professor of engineering practice in the Schools of Electrical and Computer Engineering and (by courtesy) Engineering Education and Director of the Vertically Integrated Projects (VIP) Program at Purdue University. She holds a B.S.E.E., M.S.E.E., and Ph.D. in Engineering Education, all from Purdue. Prior to this she was Co-Director of the EPICS Program at Purdue where she was responsible for developing curriculum and assessment tools and overseeing the research efforts within EPICS. Her research interests include the professional formation of engineers, diversity, inclusion, and equity in engineering, human-centered design, engineering ethics, and leadership.

Statistical Analysis and Report on Scale Validation Results for the Engineering Ethical Reasoning Instrument (EERI)

Abstract

As evidenced by the ABET criteria and numerous publications, the growing need to foster ethical awareness and judgment in engineering students is pronounced. Despite this, the ability to definitively show accreditation boards, such as ABET, that good work is being done is scarcely achievable since the most effective methods of evaluation are too time consuming. In an effort to standardize at least some means by which ethical reasoning can be measured in engineering students, a team researchers developed the Engineering Ethical Reasoning Instrument (EERI) [1]. This instrument was based on a second iteration of the Kohlbergian Defining Issues Test (DIT2). The EERI was designed to be an adaptation of the DIT2, but with scenarios contextually relevant to engineering students. Ideally, the EERI is intended to measure the degree to which participants reason "post-conventionally," described in neo-Kohlbergian theory as a belief that "moral obligations are to be based on shared ideals, are fully reciprocal, and are open to scrutiny (i.e. subject to tests of logical consistency, experience of the community and coherence with accepted practice)" [2]. Since the EERI was developed and first presented at ASEE in 2013, there have been ten additional papers presented at ASEE conferences alone which have either been directly about the EERI or have *used* the EERI as an instrument in their methodology. Now, after several years of being administered at Purdue and other institutions, there are 2000+ total responses. This paper builds on previous validation work to report formal analysis of the aggregated descriptive statistics from a sample of the total population of participants. The preliminary results from scale validation through *partial* confirmatory factor analysis of the EERI are presented and discussed.

Introduction

In 2000 the Accreditation Board for Engineering and Technology (ABET) released its new set of criteria for the turn of the century [3]. Within the criteria were three new and explicit references to ethics as an expected part of engineering curriculum. Following this there was a surge in publications regarding ethics education at the annual conferences for the American Society for Engineering Education (ASEE). A common concern which kept surfacing were questions regarding how to measure the effectiveness of new curricular efforts geared towards ethics. As noted by a team at Purdue [4], one of the only reliable scales for measuring ethical reasoning mechanisms was the current version of the Defining Issues Test (DIT2). The team noted that, although the DIT2 was a validated measure for the general population, it may not be sufficient to measure engineering-specific aspects of moral judgment [5]. In response to this criticism of the DIT2's appropriateness within engineering education, they developed the Engineering Ethics Reasoning Instrument (EERI), which was first published in the proceedings of the 2013 Annual Conference for ASEE [4].

The DIT2 is based upon neo-Kohlbergian theory which suggests that there are three primary modes of thinking, or schema, that influence decision making priorities regarding ethical dilemmas [2]. These three schemas are: 1) *pre-conventional*—the degree to which decisions are based on reference to the self and how the decision will affect the self (i.e. oriented towards self-preservation); 2) *conventional*—the degree to which one prioritizes rules, conventions, and societal

expectations when making decisions within ethical dilemmas; and 3) *post-conventional*—the degree to which one is willing to question the purpose of rules, normative behavior, societal expectations, or self-oriented consequences when making ethical decisions, preferring ideals that could benefit society as a whole while being logically coherent.

The EERI was modeled directly after the DIT2 [6], but with modified scenarios and items. Both scales contain six separate scenarios that are designed to have ethically ambiguous and value competing components. A participant reads a scenario and is then presented with twelve unique items regarding considerations that could affect how an individual might respond to the ethical dilemma. The participant is then asked to rate each of the items on a five-point Likert-scale in reference to how important they think the item is when considering how to respond to the ethical dilemma. After *rating* each of the items, the participants are then asked to *rank* the top four items which they considered to be most important. This process was carried out for all six scenarios.

Since being introduced to the engineering education community, the EERI has been administered many times and referenced in at least ten papers published through the ASEE conference proceedings [1] [7] [8] [9] [10] [11] [12] [13] [14] [15]. The EERI was referenced or used three times just last year within ASEE proceedings, one of which was purposed towards the validation of another instrument [10]. The fact that the EERI is being used within validation efforts of other instruments highlights the need to make sure that the EERI itself is validated. The EERI currently has over 2000 responses in total. Although initial validation evidence was collected [1], it is clear that additional evidence is needed to inform both current use and future improvements to the instrument [16]. Additionally, the EERI has primarily been used as an additional measure to provide depth alongside ethics education interventions to stimulate added thinking about reasoning mechanisms. Although some of the research noted above looks into how different groups of people respond differently to the EERI, it is not yet clear whether the EERI is reliably measuring something at all, or what exactly it is measuring even if shown to be valid. This gap leads to the purpose of this present paper—making progress towards achieving validation thresholds for the EERI as a scale. Once this is done, other gaps can begin to be explored.

Methodology

Partial Confirmatory Factor Analysis

The EERI has six scenarios, each with twelve items. Each item is designed to map to one of the three Neo-Kohlbergian schemas (the latent variables). The first schema is "pre-conventional", also referred to as schema-23 (based on two of the six sub-categories from Kohlberg's framework). The second schema is "conventional," also referred to as schema-4. The last schema is "post-conventional," also referred to as schema-56. Based on this theoretical structure, a preliminary analysis of item correlations was conducted to calibrate an understanding of how they were interacting (correlation matrix available upon request). This step was primarily geared towards identifying whether the intended item interactions were manifested at a surface level. Based on that examination, it was concluded that a more robust method of analysis would be necessary. A *Partial* Confirmatory Factor Analysis (PCFA) [17] [18] was chosen as the ideal starting point because it acts as a bridge between *Exploratory* Factor Analysis (EFA) and full *Confirmatory* Factor Analysis (CFA) [19]. If the number of intended factors is known (at least

theoretically, as is the case with the EERI), a PCFA can allow for unintended factors to be found while also running the necessary diagnostics for determining if a full CFA will be more likely to succeed. An additional benefit to PCFA is that it folds exploratory factor analysis into the process.

For factor analysis to be feasible and more reliable, especially when there are many items and more than two factors, large sample sizes are needed. This analysis uses a dataset of responses from Purdue students enrolled in the EPICS service learning program [20]. In the EPICS program, teams of students partner with community organizations to address real design challenges. The dataset (after cleaning) has accumulated 1178 responses over the course of several semesters.

Although the N2 scoring index (developed by Rest et al [25]) used is calculated based on item rating results *and* item ranking results, the structure of item *ranking* responses does not lend itself to factor analysis. Thus, the factor analysis will only be conducted on item *rating* results. Despite this, it is reasonable to suggest that if a consistent set of latent factors are identified as being represented by the items, the items then chosen for ranking—being the same items used for rating—will also represent the identified latent factors. The obvious limitation being that even if three factors are consistently identified, there is no guarantee that the factors being represented by the items are the actual target factors (pre-conventional, conventional, and post-conventional).

The software used to conduct the factor analysis was SPSS. The analysis went through three separate refinement cycles wherein problematic items were identified and removed to determine which sets of items most effectively represented the intended latent variables (pre-conventional, conventional, and post-conventional). The PCFA was conducted at the scenario level *and* at the aggregate level.

For each cycle of refinement, although three factors were manually selected for extraction, scree plots and item eigenvalues were surveyed to determine if there were divergences from the expected number of factors. During each of these iterations, for each scenario and the aggregate, the following results were collected: the Kaiser-Meyer-Olkin (K-M-O) measure of sampling adequacy; Bartlett's Test of Sphericity (chi-square$_{null}$, degrees of freedom, and significance); and Goodness-of-fit (chi-square$_{implied}$, degrees of freedom, and significance). The Bartlett's Test of Sphericity and Goodness-of-fit parameters (along with sample size) were then used to calculate four of the major CFA construct fit indices [18]:

- Normed-Fit Index (NFI) [21]: $NFI = \frac{\left(\chi^2_{Null} - \chi^2_{Implied}\right)}{\chi^2_{Null}}$
- Tucker Lewis Index (TLI) [22]: $TLI = \frac{\left(\chi^2_{Null}/df_{Null}\right) - \left(\chi^2_{Implied}/df_{Implied}\right)}{\left[\left(\chi^2_{Null}/df_{Null}\right) - 1\right]}$
- Comparative Fit Index (CFI) [23]: $CFI = 1 - \frac{\left(\chi^2_{Implied} - df_{Implied}\right)}{\left(\chi^2_{Null} - df_{Null}\right)}$
- the Root Mean Square of Approximation (RMSEA) [24]: $RMSEA = \sqrt{\frac{\chi^2_{Implied} - df_{Implied}}{(N-1) * df_{Implied}}}$

Where N is sample size, χ^2_{Null} is the chi-square result from the Bartlett test of Sphericity, $\chi^2_{Implied}$ is the chi-square results from the Goodness-of-fit test, df_{Null} is the degrees of freedom from

Bartlett's test, and $df_{Implied}$ is the degrees of freedom from the Goodness-of-fit test. The results of the fit indices are evaluated as follows: the NFI, TLI, and CFI are considered good if >0.95, whereas the RMSEA is "okay" if <0.08, and good if <0.06. If the majority of the fit indices meet their thresholds, then there is a reasonable chance that the construct is well represented by the scale.

During each iteration of the PCFA, the Pattern Matrix was used for determining which items loaded to which factors. If it was apparent that an item was loading on the wrong factor, or if an item was loading too weakly (<0.3) the item was removed from the pool and the next round of analysis was initiated. This took place (generally) in three cycles for each scenario (except for scenario 1).

Results

Partial Confirmatory Factor Analysis

Results from all rounds of analysis refinement for each scenario are summarized below. The referenced index values, as well as corresponding Bartlett Test of Sphericity and Goodness-of-fit results, are all detailed in summary tables below each scenario section. These tables also include, for each round, how many factors were naturally extracted by factor analysis and reference to the specific items that were removed during the process and after which round they were removed. Items are coded as follows: the first value is the schema (i.e., pre-conventional, conventional, and post-conventional), the second value indicates the scenario (i.e. s1 for scenario 1), and the last value indicates which of the items was removed.

Scenario 1

Scenario one's factor analysis successfully extracted three factors naturally on the first round and each item was correctly loaded onto a factor with the appropriate items. Although this was the only scenario that resulted in loadings perfectly consistent with theory and scale design, the fit indices were the weakest (only two above threshold). Using the Factor Matrix instead of the Pattern Matrix showed that one of the *conventional* items was cross loaded between *conventional* and *pre-conventional*. This item was dropped, resulting in another *conventional* item dropping below the strength threshold of 0.3. With the second *conventional* item removed, final analysis had all items loaded correctly and a third fit index going above its acceptable threshold.

Table 1, Statistical results of PCFA analysis for scenario 1.

Scenario	Analysis Round	Natural Factors		Bartlett's Test of Sphericity	Goodness-of-fit Test	K-M-O	Items Removed
1	1	3	χ^2	2687.123	154.321	0.728	1 item removed
			df	66	33		
			Sig	0.000	0.000		
	2	3	χ^2	2387.074	95.514	0.717	1 item removed
			df	55	25		

			Sig	0.000	0.000		
	3	3	χ^2	2219.208	71.978	0.707	
			df	45	18		
			Sig	0.000	0.000		

Scenario 2

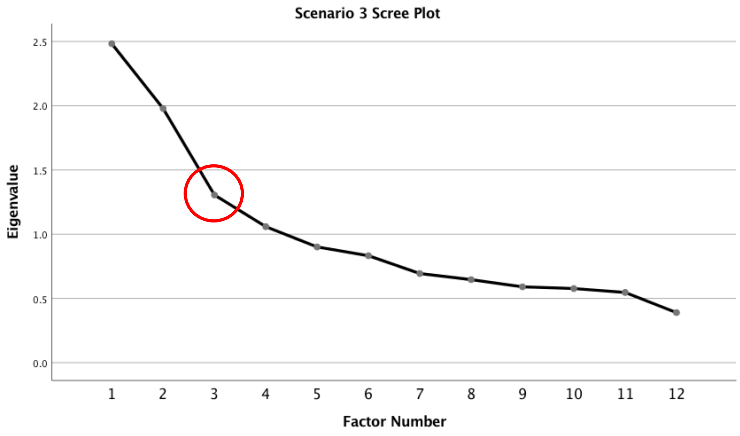
Scenario two's factor analysis successfully extracted three factors naturally on the first round, but there was an item which was incorrectly loaded (a *post-conventional* item loaded with the *conventional* items). After removing this item, all remaining items were correctly loaded, but one of the schema-4 items dropped below a loading strength of 0.3 and was removed before a final analysis. The final analysis resulted in correctly loaded items and three fit indices above threshold.

Table 2, Statistical results of PCFA analysis for scenario 2.

Scenario	**Analysis Round**	**Natural Factors**		**Bartlett's Test of Sphericity**	**Goodness-of-fit Test**	**K-M-O**	**Items Removed**
2	1	3	χ^2	2894.233	218.224	0.734	1 item removed
			df	66	33		
			Sig	0.000	0.000		
	2	3	χ^2	2686.291	199.331	0.728	1 item removed
			df	55	25		
			Sig	0.000	0.000		
	3	3	χ^2	2377.840	92.094	0.715	
			df	45	25		
			Sig	0.000	0.000		

Scenario 3

Scenario three's factor analysis naturally produced 4 factors. This was because the threshold was set to 1 eigenvalue. Upon analyzing the scree plot (shown in figure 1), it was clear that a) the fourth factor was only just above an eigenvalue of 1, and b) there was a stronger inflection point at the third factor. Thus, it was concluded that 3 factors were appropriate.



Upon analysis of the Pattern Matrix, two of the *conventional* items were shown to be below a loading strength of 0.3 and were removed. The second analysis improved greatly but still had one rogue *conventional* item which was loaded with the *pre-conventional* items. It was removed, and the final analysis showed good loading fit for the remaining items and all four fit indices were above threshold.

Table 3, Statistical results of PCFA analysis for scenario 3.

Scenario	Analysis Round	Natural Factors		Bartlett's Test of Sphericity	Goodness-of-fit Test	K-M-O	Items Removed
3	1	4	χ^2	2109.649	159.737	0.684	2 item removed
			df	66	33		
			Sig	0.000	0.000		
	2	3	χ^2	1844.211	57.081	0.671	1 item removed
			df	45	18		
			Sig	0.000	0.000		
	3	3	χ^2	1680.246	31.405	0.654	
			df	36	12		
			Sig	0.000	0.000		

Scenario 4

Although five factors were extracted naturally, running again while restricting to only five factors resulted in two of the factors dropping below an eigenvalue of 1, so three factors were justified. Items which were above a loading strength of 0.3 all fit with theoretical expectation on the first round, so the two lowest items—one from *pre-conventional* and one from *conventional* —were removed. The second round of analysis only had one remaining *post-conventional* item which was below 0.3 and was removed. The final round of analysis had good results, except one of the *pre-conventional* items is weakly loaded with *conventional* items (at a strength of 0.323) in addition to a stronger loading (0.520) with the remaining schema-23 items. Final round analysis also yielded three fit indices above threshold.

Table 4, Statistical results of PCFA analysis for scenario 4.

Scenario	Analysis Round	Natural Factors		Bartlett's Test of Sphericity	Goodness-of-fit Test	K-M-O	Items Removed
4	1	5	χ^2	2232.400	305.388	0.700	2 items removed
			df	66	33		
			Sig	0.000	0.000		
	2	3	χ^2	1687.303	68.810	0.677	1 item removed
			df	45	18		
			Sig	0.000	0.000		
	3	3	χ^2	1530.058	44.133	0.657	
			df	36	12		
			Sig	0.000	0.000		

Scenario 5

Scenario five's factor analysis cleanly extracted only three factors. First round analysis of the Pattern Matrix looked good, except that there was a *conventional* item loaded with the *post-conventional* items and a *post-conventional* item loaded with the other *conventional* items. These two items were removed. The second round of analysis a *pre-conventional* item was split loaded with *conventional* and was removed. The final round of analysis had good loading with remaining items and all four fit indices were above threshold.

Table 5, Statistical results of PCFA analysis for scenario 5.

Scenario	Analysis Round	Natural Factors		Bartlett's Test of Sphericity	Goodness-of-fit Test	K-M-O	Items Removed
5	1	3	χ^2	2228.127	178.373	0.747	2 items removed
			df	66	33		
			Sig	0.000	0.000		
	2	3	χ^2	1775.411	88.749	0.726	1 item removed
			df	45	18		
			Sig	0.000	0.000		
	3	3	χ^2	1427.591	28.331	0.708	
			df	36	12		
			Sig	0.000	0.000		

Scenario 6

Scenario six's factor analysis extracted 4 factors naturally when all items were included. Scree plot analysis did not yield any clear justification for reducing factors to only three. It was decided to force a three-factor extraction and clean erroneous items until a natural three factor extraction was achieved. First round analysis of the Pattern Matrix showed relatively clean loading except two *post-conventional* items below loading strength of 0.3, which were removed. There was an additional *pre-conventional* item which was strongly loaded with *post-conventional*. The strength of the loading (stronger than either of the remaining *post-conventional* items) suggested that it was legitimate, suggesting that the item may in fact need to be changed to a *post-conventional* item. Second round analysis resulted in one *conventional* item below loading strength of 0.3 and was removed. Final round analysis resulted in good loading (assuming that the rogue *pre-conventional* item can be assumed now as a *post-conventional* item).

Table 6, Statistical results of PCFA analysis for scenario 6.

Scenario	Analysis Round	Natural Factors		Bartlett's Test of Sphericity	Goodness-of-fit Test	K-M-O	Items Removed
6	1	4	χ^2	2198.483	73.133	0.790	2 items removed
			df	66	24		
			Sig	0.000	0.000		
	2	3	χ^2	1884.542	59.135	0.776	1 item removed
			df	45	18		

			Sig	0.000	0.000		
	3	3	χ^2	1797.045	46.415	0.778	
			df	36	12		
			Sig	0.000	0.000		

Index performance

Prior to removing any items, when running factor analysis with the current form of the EERI, index performance is generally poor. As is shown in Table 7, only three of the scenarios had at least one index which performed within acceptable thresholds. None of them met the preference of having at least three indices which performed within acceptable thresholds.

Table 7, Index performance results for all scenarios prior to any item removal. *"*" Near threshold, "**" above threshold, unacceptable where there are no asterisks*

First Round Analysis Results				
Scenario/Index	**NFI (>0.950)**	**CFI (>0.950)**	**TLI (>0.950)**	**RMSEA (<0.06 or <0.08)**
Scenario 1 Items	0.943 *	0.954 **	0.908	0.0558 **
Scenario 2 Items	0.925 *	0.935 *	0.869	0.0691 *
Scenario 3 Items	0.924 *	0.938 *	0.876	0.0571 **
Scenario 4 Items	0.863	0.874	0.749	0.0837
Scenario 5 Items	0.920 *	0.933 *	0.866	0.0612 *
Scenario 6 Items	0.940 *	0.954 **	0.908	0.0503 **
Combined Scenario Items	0.621	0.685	0.656	0.0497 **

After refinement of item selection, a much better index performance was achieved. As can be seen in Table 8, all scenarios now meet at least the minimum expectation of a majority of indices having performance within acceptable thresholds. Even the TLI index performance was relatively close to the threshold of >0.95, with an average deficiency of only 0.016. The only remaining set of indices which completely failed regardless of item selection are those associated with factor analysis of all items from the entire EERI in aggregate.

Table 8, Index performance results for all scenarios after completing analysis iterations and removing select items. *"*" Near threshold, "**" above threshold, unacceptable where there are no asterisks*

Final Round Analysis Results				
Scenario/Index	**NFI (>0.950)**	**CFI (>0.950)**	**TLI (>0.950)**	**RMSEA (<0.06 *or* <0.08)**
Scenario 1 Items	0.968 **	0.975 **	0.938 *	0.0505 **
Scenario 2 Items	0.961 **	0.968 **	0.921 *	0.0591 **
Scenario 3 Items	0.981 **	0.988 **	0.965 **	0.0371 **
Scenario 4 Items	0.971 **	0.979 **	0.936 *	0.0477 **
Scenario 5 Items	0.980 **	0.988 **	0.965 **	0.0340 **
Scenario 6 Items	0.974 **	0.980 **	0.941 *	0.0494 **

Combined Scenario Items	0.667	0.717	0.684	0.0543 **

Removed Items

Throughout the iterations of item selection refinement, a total of sixteen items were identified as poorly fitted and removed. Two items were removed from the *pre-conventional* pool; nine items were removed from the *conventional* pool; and five items were removed from the *post-conventional* pool. This left 56 items in the scale from the original 72.

Discussion

The PCFA demonstrated that the EERI, in its current form, does not seem to adequately target the intended latent construct. Although the desired construct seems likely to be achievable through the elimination of certain items, this introduces new problems for the EERI. First, with the elimination of 16 of the 72 items, the structure of the EERI no longer fully matches that of its model, the DIT2. The principle structure of the EERI would still be consonant with the DIT2, but a 22% reduction in the number of items introduces some strain to direct comparisons. Despite this, we believe that the principle structure regarding how items are presented is more dominant than *how many* items are presented.

The second problem with item elimination is the unbalanced way that the items were removed. *Conventional* items were the most commonly removed items, making up 56% of those eliminated. Next were *post-conventional* items at about 31%, and finally only about 13% were *pre-conventional*. Not only was the distribution of removed items unequal, but the distribution of which scenarios the items were removed from was also unequal. It is not clear yet whether this is problematic, and further analysis will need to be conducted.

Lastly, and most importantly, the elimination of items may nullify the usefulness of the N2 index for obtaining a final score, which is likely dependent upon a relatively equal number of *pre-conventional* and *post-conventional* items. That being said, it is not clear yet how sensitive the N2 index is to inequality between these schemas. The DIT2 and the EERI (with all 72 items) already have an imbalance in this regard, having only 23 *pre-conventional* items compared to 24 *post-conventional* items. It is possible that the N2 has been compromised in some way due to the larger disparity that now exists.

Moving forward, these results lead to two options. One option would be an attempt to replace the missing items with new items. Although this would solve the imbalance issues noted above and bring the EERI back into consonance with the DIT2, it would be a lengthy and extensive process. Creating new items would require a new set of data to be generated and yet another round of involved factor analysis before finding out if the new items work. An alternative option is simply to remove to 16 items and leave the EERI in its new form. In this case, more analysis would still need to be conducted to complete a full Confirmatory Factor Analysis, but that can be conducted with existing data (since this analysis was conducted using less than half of the total data available). If it can be shown that the target construct really is represented by the EERI, the next step would be to either re-validate the now potentially compromised N2 index, revert to using the P index, or try and generate a new index that factors in the imbalance.

One method for validating an index for this scale would be to mirror what Rest, Narvaez, and Bebeau [25] did when originally creating and validating the N2 index. Their approach was to use scores generated by experimental indices and test for sensitivity to the following criteria—also used to validate the original DIT [26]:

- Sensitivity to educational intervention
- Sensitivity to differentiated educational groups
- Sensitivity to longitudinal trends
- Correlations with moral comprehension
- Links to behavior
    - Civil libertarian attitudes
    - Pro-social behavior

As administrations of the EERI continue, it may be necessary to start including validated methods of measuring the above listed factors to validate the N2 index or another index which may be designed. Detecting these sensitivities will also alleviate some of the uncertainty about whether the three factors found within each scenario are the target factors (pre-conventional, conventional, and post-conventional)

Limitations

The current analysis was conducted primarily with only students from Purdue and within the context of one course. This is not a broad enough sample to make any sweeping conclusions. There was a moderate selection of students represented from sophomore, junior, and senior years, but the majority of the students were freshman. A further round of analysis will be conducted once the additional 1000+ set of responses—many of which came from other courses and institutions—has been cleaned and prepared for factor analysis. Considering that the fit indices do not hold when the items are analyzed in aggregate, there is a chance that different types of factors are being represented by each scenario. It is also possible that the breakdown of cohesion at the aggregate level is due to the changing context of each scenario. Future research will need to explore how to resolve this uncertainty.

Conclusion

With the EERI now several years old and having been used many times with well over 2000 responses, a formal analysis of construct validity was needed. Using a subset of the data collected from students in the EPICS service-learning program at Purdue, a *Partial* Confirmatory Factor Analysis was conducted. This analysis resulted in the identification of 16 (out of 72) items which were problematic. When these 16 items were included in factor analysis, construct fit indices performed very poorly. When these items were removed, the construct fit indices had very promising results, suggesting that a full Confirmatory Factor Analysis would be worth pursuing and likely to succeed. At this point, two primary pathways could be taken for future work regarding the EERI: 1) new items are created to replace the ones that need to be removed; or 2) the problematic items are removed, no items replace them, and the scoring index is re-validated or re-created.

References

[1] Zhu, Q., Zoltowski, C., Feister, M., Buzzanell, P., Oakes, W., and Mead, A. (2014). The development of an instrument for assessing individual ethical decisionmaking in project-based design teams: integrating quantitative and qualitative methods. In: *American Society for Engineering Education Annual Conference*. Indianapolis, IN.

[2] Rest, J., Narvaez, D., Thoma, S., and Bebeau, M. (2000). A neo-Kohlbergian approach to morality research. *Journal of Moral Education, 29*(4), pp. 381-395.

[3] Frey, W., Sanchez, H., and Cruz, J. A. (2002). Ethics across the curriculum: An effective response to ABET 2000. In: *American Society for Engineering Education Annual Conference*. Montreal, CA.

[4] Zoltowski, C., Buzzanell, P., and Oakes, W. (2013). Utilizing an engineering ethical reasoning instrument in the curriculum. In: *American Society for Engineering Education Annual Conference*. Atlanta, GA.

[5] Bebeau, M. (2002). The Defining Issues Test and the Four Component Model: Contributions to professional education. *Journal of Moral Education*, 31(3), pp.271-295.

[6] Rest, J., Narvaez, D., Thoma, S. and Bebeau, M. (1999). DIT2: Devising and testing a revised instrument of moral judgment. *Journal of Educational Psychology*, 91(4), pp.644-659.

[7] Bielefeldt, A., Canney, N., Swan, C. and Knight, D. (2019). Efficacy of macroethics education in engineering. In: *American Society for Engineering Education Annual Conference*.

[8] Bielefeldt, A., Polmer, M., Kniht, D. and Canney, N. (2017). Incorporation of ethics and societal impact issues into first-year engineering course: Results of a national survey. In: *American Society for Engineering Education Annual Conference*. Columbus, OH.

[9] Burkey, D. and Young, M. (2017). Work-in-progress: A 'cards against humanity'-style game for increasing engineering students' awareness of ethical issues in the profession. In: *American Society for Engineering Education Annual Conference*. Columbus, OH.

[10] Butler, B., Anastasio, D., Burkey, D., Cooper, M. and Bodner, C. (2018). Work in progress: Content validation of an engineering process safety decision-making instrument (EPSRI). In: *American Society for Engineering Education Annual Conference.* Salt Lake City, UT.

[11] Cimino, R., and Steiner, S. (2018). Effectiveness of ethical interventions in a first-year engineering course: A pilot study. In: *American Society for Engineering Education Annual Conference*. Salt Lake City, UT.

[12] Ghorbani, M., Maciejewski, A., Siller, T., Chong, E., Omur-Ozbek, P., and Atadero, R. (2018). Incorporating ethics education into an electrical and computer engineering undergraduate program. In: *American Society for Engineering Education Annual Conference*. Salt Lake City, UT.

[13] Hess, J., Kisselburgh, L., Zoltowski, C., and Brightman, A. (2016). The development of ethical reasoning: A comparison of online versus hybrid delivery modes of ethics instruction. In: *American Society for Engineering Education Annual Conference*. New Orleans, LA.

[14] Kisselburgh, L., Zoltowski, C., Beever, J., Hess, J., Iliadis, A., and Brightman, A. (2014). Effectively engaging engineers in ethical reasoning about emerging technologies: A cyber-enabled framework of scaffolded, integrated, and reflexive analysis of cases. In: *American Society for Engineering Education Annual Conference*. Indianapolis, IN.

[15] Kisselburgh, L., Hess, J., Zoltowski, C., Beever, J., and Brightman, A. (2016). Assessing a scaffolded, interactive, and reflective analysis framework for developing ethical reasoning

in engineering students. In: *American Society for Engineering Education Annual Conference*. New Orleans, LA.
[16] Douglas, K. and Purzer, Ş. (2015). Validity: Meaning and Relevancy in Assessment for Engineering Education Research. *Journal of Engineering Education*, 104(2), pp.108-118.
[17] Bollen, K. (2014). *Structural Equations with Latent Variables*. New York, NY: John Wiley & Sons.
[18] Gignac, G. (2009). Partial Confirmatory Factor Analysis: Described and Illustrated on the NEO–PI–R. *Journal of Personality Assessment*, 91(1), pp.40-47.
[19] Bryant, F. and Yarnold, P. (1995). Principal-component analysis and exploratory and confirmatory factor analysis. In: G. Grimm and P. Yarnold, ed., *Reading and Understanding multivariate statistics*. Washington, DC: American Psychological Association, pp.99-136.
[20] Zoltowski, C. B. & Oakes, W. C. (2014). Learning by doing: Reflections of the EPICS Program, *International Journal for Service Learning in Engineering*, Special Edition Fall 2014, 1-32.
[21] Bentler, P. and Bonett, D. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88(3), pp.588-606.
[22] Tucker, L., and Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, pp. 1–10.
[23] Bentler, P. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107(2), pp.238-246.
[24] Browne, M. and Cudek, R. (1993). Alternative ways of assessing model fit. In: K. Bollen and J. Long, ed., *Testing structural equation models*. Newbury Park, CA: Sage, pp.136-162.
[25] Rest, J., Thoma, S., Narvaez, D. and Bebeau, M. (1997). Alchemy and beyond: Indexing the Defining Issues Test. *Journal of Educational Psychology*, 89(3), pp.498-507.
[26] Rest, J., Thoma, S. and Edwards, L. (1997). Designing and validating a measure of moral judgment: Stage preference and stage consistency approaches. *Journal of Educational Psychology*, 89(1), pp.5-28.