Initial Validation of Indirect Assessments tools for Connections and Creating Value for Entrepreneurial Minded Learning

Tyler James Stump, The Ohio State University

Tyler Stump is a second-year Ph.D. student in the Department of Engineering Education at The Ohio State University. Tyler received his B.S. in Biosystems Engineering at Michigan State University in 2022 and received his M.S. from Michigan State University in 2023. His engineering education interests include critical quantitative research methods, computing education, and assessment validation.

H. Schwab, The Ohio State University

H. Schwab is pursuing a B.S. in Chemical Engineering with a minor in Humanitarian Engineering at The Ohio State University. Involvement includes working as an Undergraduate Research Associate and Lead Undergraduate Teaching Associate for the Fundamentals of Engineering Program within the Department of Engineering Education. Research interests focus on sense of belonging, concept mapping, instrument content validation, and metacognition.

Sydney Cooper, The Ohio State University

Sydney Cooper is pursuing her B.S. in Biomedical Engineering at The Ohio State University. She is involved in the Department of Engineering Education as an Undergraduate Research Associate. Her research interests include inclusion in engineering.

Dr. Krista M Kecskemety, The Ohio State University

Krista Kecskemety is an Associate Professor in the Department of Engineering Education at The Ohio State University and the co-Director of the Fundamentals of Engineering Programs. Krista received her B.S. in Aerospace Engineering at The Ohio State University in 2006 and received her M.S. from Ohio State in 2007. In 2012, Krista completed her Ph.D. in Aerospace Engineering at Ohio State. Her engineering education research interests include investigating first-year engineering student experiences, faculty experiences, and the research to practice cycle within first-year engineering.

Initial Validation of Indirect Assessments Tools for Connections and Creating Value for Entrepreneurial Minded Learning

Abstract

Entrepreneurial Minded Learning is a commonality shared amongst educators of the Kern Entrepreneurial Engineering Network (KEEN) that embraces entrepreneurial pedagogy infused in the engineering classroom through the 3Cs: Curiosity, Connections, and Creating Value. However, there currently are very limited assessments tools available to educators hoping to adopt EML in their classroom. This gap in assessment tools limits the degree to which student learning gains and educational intervention effectiveness can be measured in the classroom both directly and indirectly. As a result, this research paper focuses on the initial validation for indirect assessments for Connections & Creating Value, the final two missing assessment tools for the assessment bundle sought after.

The initial validation analysis was anchored in identifying face validity and content validity of the instrument. Expert judgements from reviewers with expertise in EML and/or instrument construction were collected for assessment items for each item aligned to the two indirect assessments. Validation evidence interrogated both the instruments (scales) and the assessment question (items). Face validity was assessed using statistical analyses including Item Face Validity Index and Average Scale Face Validity Index. Content validity, reflecting expert opinions on the instrument's relevance, was evaluated using metrics such as the Item Content Validity Index. Fleiss' κ was used to measure expert agreement, guiding item reduction to enhance the instrument's validity before further validation studies.

Expert reviewers assessed the face validity, relevance, clarity, and essentiality of items within the indirect assessments. Based on their insights, it was determined that the variability of how KEEN Experts reviewed the items was beyond initially hypothesized. Thus, item reduction could not solely be informed by a quantitative analysis. Future research will integrate the open-ended comments reviewers left to scope in why this high variability occurred and what item elements were a mechanism for this high variability. This initial investigation into the validation of these instruments supports a larger endeavor to advance assessment tools for entrepreneurial engineering education. Equipping engineering educators with adequate and nuanced assessments tools could enhance the ways in which best EML practices in the classroom are evaluated with hopes of ultimately improving EML skillsets for engineering students entering professional practice.

1. Introduction

Engineers are positioned to be impactful contributors to solving modern global problems such as climate change, food shortages, and sustainable energy [1]. These complex modern challenges often are ill-structured and require engineers to apply technical skills such as computational modeling to solve these problems [2]. Oftentimes, these complex global problems are embedded within social systems which need to be accounted for when developing effective engineering solutions [3]. This interplay between social realities and technical skill application leads to *sociotechnical engineering* solutions. Sociotechnical engineering refers to the blended necessity

of social responsibility embedded within engineering ways of doing [3]. The conceptualization of social responsibility with technical skills was first introduced by Mackenzie and Wajcman (1985) in their publication *"The Social Shaping of Technology."* The researchers argued that technology is influenced by social contexts and requires those developing and applying technology must deeply consider the social factors shaping the technological use and/or development [4]. Though the researchers did not specifically anchor this notion with the engineering profession, it does deeply impact those engaging with technology. Carl Mitcham in 1994 anchored the consideration of social responsibility into engineering practice and profession. By framing engineering as a social enterprise, he argued that the integration of social considerations is inherent to engineering practice is required to better equipped modern-day engineers with capacity to solve not only technical challenges but social and ethical ones as well [5]. Mitcham's introduction of a *sociotechnical engineer* was catalyzed later by the National Academy of Engineering's 2005 report, *"Educating the Engineer of 2020: Adapting Engineering Education to the New Century"* that identified that the education of young engineers provided an opportunity to build students capacity for blending these social and technical skills [2]. Thus, the catalyzation of considering emerging contexts capable of support student learning for sociotechnical engineering development became a priority for engineering education practitioners and researchers [6], [7].

One emerging context with the ability to support the development of sociotechnical skills in engineering students is through *entrepreneurship* blended within engineering education. Entrepreneurial education is not simply preparing students to start a business, rather "…to develop to the students the knowledge, skills and competencies which will help them to engage in a more enterprising, innovative and flexible manner in the changing workplace environment from today" [8]. When centered in engineering curriculum, entrepreneurship allows for students to engage with skills such as empathy, collaboration, and creativity [9], [10], [11]. The Kern Entrepreneurial Engineering Network (KEEN) is a partnership of more than 55 colleges and universities across the United States that work to support engineering educators with tools, assessments, and resources in developing engineering student's entrepreneurial mindset [12]. More specifically, "The Entrepreneurial Mindset (EM) is a set of attitudes, dispositions, habits, and behaviors that shape a unique approach to problem solving, innovation and value creation," [13]. The KEEN network works to establish the research to practice cycle in engineering in informing ways to heighten engineering students engagement with empathy, creativity, resiliency, flexibility, and collaborative abilities through entrepreneurial education in engineering classroom pedagogical approaches [14], [15], [16].

The KEEN network has adopted a lens of these pedagogical approaches through "the 3 Cs" referring to *Curiosity, Connections,* and *Creating Value*. KEEN educators use the 3Cs to expand students' sociotechnical skills through entrepreneurial constructs. This tridimensional approach allows educators to connect engineering topics, such as thermodynamics or mass balance modeling, by probing students to be curious about the world, to make connections between knowledge, and to identify opportunities to create value for the world. Research on the 3Cs has been investigated for quite some time with a particular focus on topics such as educational interventions and mindset in contexts such as the first-year engineering programs, the mid-years, and the senior capstone courses. Despite the ubiquitous research efforts looking to expand pedagogical approaches for the 3Cs, there remains a large gap in tools available for educators to

measure student learning gains and pedagogical intervention effectiveness. This project is part of a larger research effort to support the development of direct and indirect assessments for each of the 3Cs (*Curiosity, Connections & Creating Value)*. This "assessment bundle" will allow for educators to have malleable measurement instruments capable of supporting a wide range of variety in the ways in which engineering educators may choose to integrate the 3Cs into their courses. As of now, all the direct assessments and the indirect assessment for curiosity have been developed and validated – leaving only the indirect assessments for Connections & Creating Value to complete the "assessment bundle". Thus, this research paper will focus on the initial validation of the indirect assessments for Connections & Creating Value.

2. Background

Educational assessment instrument validation's purpose and definition is a contentious point amongst psychometrics instrument developers. Validation generally refers to the interpretations of assessing data of an instrument that supports the plausibility that what is intended to be measured is being measured [17]. However, an instrument itself cannot be validated, despite ubiquitous language that encapsulates this misconception. This tension between psychometrics instrument developers lies in this notion; however, context matters, and as such it becomes impossible to gather adequate validity evidence conducive to supporting an instrument capable of being "valid" in every context [18], [19], [20]. This research paper anchors in the notion that validation is an ongoing process of various sources of evidence to showcase whether what is intended to be measured is truly being measured, specific to context [17], [20].

Validation evidence thus must be investigated through multiple forms to provide sufficient means of validation within contexts [21]. Validation takes numerous forms and interrogates the instruments development through four primary forms: face validity, content validity, construct validity, and criterion validity. This initial validation evidence analysis focuses on face and content validity as the initial components to the validation investigation. Face validity investigates the appropriateness of assessment items that are to be included within the measurement tool, typically through the form of expert judgement. More specifically, face content asks expert reviewers to judge "at face value" the extent to which the items align with the intention of the indirect assessment's constructs [22]. Reviewers typically are probed with a yes/no binary question on whether they believe the item should be included in the scale or not. Content validity pushes this initial component of validation evidence. *Content validity* refers to the extent to which the items of an instrument are associated with the construct of interest from. That is, it reflects the extent to which experts in the field evaluate that the items do relate to the construct being measured by considering how essential, how clear, and how relevant each item is through quantitative and qualitative approaches [23]. Thus, the need to create a survey instrument and collect expert reviewer judgement data for the indirect assessments on Connection and Creating Value became the next step and is described below.

3. Methods

3.1 Data Collection

The pre-pilot study for the indirect assessments for Connections & Creating Value were both designed to collect evidence for face and content validity primarily. Face and content validity anchor on the use of expert reviewer feedback. Therefore, recruiting expert reviewers for feedback on these indirect assessments began. Recruitment efforts identified experts in two forms as either a content expert in KEEN and the 3Cs or as an instrument development expert. This aligns with best practices in allowing reviewers to interrogate both the content and the way in which the question is presented from multiple lenses [24]. A Qualtrics survey was developed to allow reviewers to evaluate each item with indirect assessments [25]. We chose to allow expert reviewers to provide feedback on either the indirect assessment for Connection or Creating Value but not both. There were three primary forms of question types utilized in the survey to support acquiring quantitative and qualitative feedback to support item reduction. The first form utilized binary responses of yes or no. Question A of the survey as seen below in Table 1 embraced this question form to provide binary data needed to evaluate face validity. Likert scales [26] are the second form embedded into the survey in which Questions B, C, and D all utilized Likert scales. Question B utilizes a 4-point Likert scale while Questions C and D utilize a 1-3 Likert scale. These three questions provide the basis to support evidence and claims of content validity. By utilizing Likert scales, we provide adequate means of data to identify expert judgement reviews suggesting item reductions or improvements. The final question form utilized in the survey are open-ended responses. Questions E and F ask reviewers more general thoughts or improvements that are offered through qualitative responses specific to each of their reviewing lenses. Questions E and F were not directly utilized within the face and content validity analysis; however, these recommendations and comments will be used in the future to support informed item reduction before determining construct validity of the indirect assessments as the next step. Table 1 below summarizes these questions, forms, and validity form it aligns with.

Table 1 - *Survey Questions adopted from* [25]

Label	Question	Question Form	Validity Construct
A	Do you think the statement measures an element of Connection/Creating Value making skills for students	Binary *Yes-No*	Face Validity
B	How relevant is this item to measuring Connection/Creating Value making skills for students?	Likert *1-4 scaling*	Content Validity
C	How clear is this item?	Likert *1-3 scaling*	
D	How essential is this item?	Likert *1-3 scaling*	
E	Recommendations for improvement of this item	Open-ended	Varies
F	Any additional comments about the Connection/Creating Value assessment or items that think are not captured?	Open ended	Varies

3.2 Data Analysis

Upon collecting expert reviewer data, quantitative analysis for both face validity and content validity evidence can be evaluated using various metrics. Face validity evidence for the instrument was evaluated using Item Face Validity Index (IFV-I), Universal Agreement Scale Validity (S-IFV/UA), and Average Scale Face Validity (S-IFV/Ave) [27]. The IFV-I indicates the percentage of raters who assign an item with clarity of 3 or 4. The S-IFV/Ave is calculated

by averaging the IFV-I scores across all items on the scale, or alternatively, the mean clarity and comprehension ratings from all raters. The proportion of clarity is determined by averaging the individual ratings provided by each rater. The S-IFV/UA refers to the proportion of items on the scale that receive clarity ratings of 3 or 4 from all raters. The Universal Agreement (UA) score is assigned a value of 1 if all raters agree on item, and 0 if there is any disagreement among raters [27]. These metrics utilize data from reviewers' response on *"Do you think…"* as binary data. The equations for these metrics for face validity can be seen below in Equations 1,2, and 3. Benchmarks to be included beyond future item reduction would require scores greater than or equal to 0.8. Below Eq(s). 1-3 describe these metrics used for face validity evidence.

IFV-I = # of Yes /Total by item Eq. 1

S-IFV/UA = #items with IFV of 1 / total # of items Eq. 2

S-IFV/Ave = Sum of Scores/Total Number of Items Eq. 3

Content validity evidence utilized the Likert-scale questions to explore validation evidence through item level metrics such as Item Content Validity Index (ICV-I) and Content Validity Ratio (CVR) as well as the scale level with Universal Scale Agreement (S-CVI/UA) and Average Scale Face Validity (S-CVI/Ave). The ICV-I represents the percentage of content experts who rate an item as relevant with a score of 3 or 4. The S-CVI/Ave is the average of the I-CVI scores across all items on the scale, or alternatively, the average of the relevance ratings provided by all experts. The proportion of relevance is calculated by averaging the relevance ratings provided by each expert. The S-CVI/UA reflects the percentage of items on the scale that receive relevance ratings of 3 or 4 from every expert. The UA score is 1 when all experts agree on an item, and 0 if there is any disagreement [28]. Benchmarks for the various items were informed by prior literature. I-CVI metrics are marked acceptable above 0.79, S-CVI/UA scores above 0.80, and S-CVI/Ave scores above 0.9 to be aligned with the goals of the indirect assessment in establishing a high-quality measurement tool [29].

Lawshe's Content Validity Ration method [30] has been widely used within psychometric instrument validation such as health studies, organizational development, and higher education [31], [32]. Content Validity Ratio (CVR) is believed by scholars to be a focused method of interrogating expert reviewer's opinion on the essentiality of an item. CVR spans values from -1 to +1 with those closer to +1 demonstrating a high agreement of essentiality across expert reviewers and those closer to -1 demonstrate disagreement in reviewers. The mathematical model for CVR can be seen below in equation 7. Given that five reviewers reviewed each indirect assessment, we adopted recommendations of benchmark minimum values for inclusion in the next iteration of the scale to be above 0.99 [32]. This investigation into validation evidence for content validity thus included this metric as a quantitative means to explore reviewers alignment in the essentiality of items within the indirect assessments for Connections & Creating Value.

$I\text{-}CVI_R$ = # of High Relevance (4) / Total by item *(Relevance)* Eq. 4a

I-CVIc = # of High Clarity (3) / Total by item *(Clarity)* Eq. 4b

$S\text{-}ICV/UA_R$ = #items with ICV-I of 1 / Total Number of Items *(Relevance)* Eq. 5a

S-ICV/UAc = #items with ICV-I of 1 / Total Number of Items *(Clarity)* Eq. 5b

$S\text{-}ICV/Ave_R$ = Sum of Scores/Total Number of Items *(Relevance)* Eq. 6a

S-ICV/Ave$_C$ = Sum of Scores/Total Number of Items *(Clarity)* Eq. 6b

CVR = (Ne – $N/2$)/(N/2) *(Essential)* Eq. 7

Ne = Number of panelists indicating "essential"

N = Total number of panelists/reviewers

A final set of analyses was conducted on the quality of expert review feedback data utilized. More specially, we looked to quantify the strength of agreement of expert reviewers scoring. To accomplish this interrater reliability, Fleiss' κ statistical analyses [33], [34] were used to quantify the strength of agreement of expert reviewers specifically for item relevance and clarity. Feiss κ considers the Item Content Validity Index (ICV-I) and the probability of chance across the agreement from expert reviewers. The equations for Probability of chance (Eq. 8) and Feiss κ (Eq. 9) can be seen below. Scholars have identified acceptable Fleiss' κ benchmarks as follows: 0.4-0.59 (fair), 0.60-0.74 (good), and above 0.74 (excellent) [33]. Below we present the findings of our face and content validity analyses.

$$P_C = \frac{N!}{A!\,(N-A)!} * 0.5^N \quad \text{Eq. 8}$$

N = number of experts

A = number of experts agreeing on high marks for specific content

$$\kappa = \frac{\mathrm{ICV\_I} - \mathrm{P_C}}{1 - \mathrm{P_C}} \quad \text{Eq. 9}$$

4. Results

As of now, eleven expert reviewer judgements have been collected via this survey in which all eleven have provided qualitative response, but only ten provided quantitative reviews. To support the necessary quantitative evidence needed to establish face and content validity, only the ten expert reviews including both qualitative and quantitative were used in this process with the additional reviewers' comments supporting additional item reduction in the future. These eleven expert judgements allow us to investigate through statistical analysis what evidence exists for face and content validation arguments.

To begin, Table 4 below describes the identified metrics for face validity evidence including each item's IFV-I, S-IFV/UA, and S-IFV/Ave. Overall, both indirect assessments showed a moderate status from the evaluation. On the scale level, both S-IFV/Ave fell beneath the benchmark of 0.8 sought after. At the item level, only one item on each of the indirect assessments fell short of the 0.8 benchmark established including CON-10 and CV-6. Thus, these items upon expert reviewers' judgement should most likely be removed in future iterations of the instrument based on heightening face validity evidence and argumentation.

Table 2 – *Face Validity Evidence*

Item	*IFV-I*	*S-IFV/UA*	*S-IFV/Ave*
CON-1	1	0.44	0.87
CON-2	0.8		
CON-3	0.8		
CON-4	1		
CON-5	1		
CON-6	1		
CON-7	0.8		
CON-8	0.8		
CON-9	1		
CON-10	0.4**		
CON-11	0.8		
CON-12	0.8		
CON-13	1		
CON-14	1		
CON-15	0.8		
CON-16	1		
CON-17	0.8		
CON-18	0.8		

Item	*IFV-I*	*S-IFV/UA*	*S-IFV/Ave*
CV-1	1.0	0.57	0.89
CV-2	0.8		
CV-3	0.8		
CV-4	1.0		
CV-5	1.0		
CV-6	0.6**		
CV-7	1.0		
CV-8	1.0		
CV-9	0.8		
CV-10	1.0		
CV-11	1.0		
CV-12	0.8		
CV-13	1.0		
CV-14	1.0		
CV-15	1.0		
CV-16	0.6**		
CV-17	1.0		
CV-18	0.8		
CV-19	0.8		
CV-20	1.0		
CV-21	0.8		

Note: ***Below Benchmark*

Content validity evidence contains much more rigorous investigation than that of face validity above govern the tri-dimensional considerations expert reviewers are asked to consider (Relevance, Clarity, and Essentiality). Table 5 showcases the results of the analysis for Connections indirect assessment while Table 6 showcases the results for Creating Value indirect assessment. Both analyses utilized item level evidence and scale level evidence for relevance, clarity, and essentiality scoring collected from the expert reviewers.

Table 3 – *Content Validity Evidence (Connections)*

Item	$I\text{-}CVI_R$	$I\text{-}CVI_C$	*CVR*	$S\text{-}CVI_R$	$S\text{-}CVI_C$	$AV\text{-}CVI_R$	$AV\text{-}CVI_C$
CON-1	0.6*	0.8	-0.27	0.06	0.22	0.48*	0.69*
CON-2	0.4**	0.8	-0.27				
CON-3	0.0**	0.0**	-1.00				
CON-4	0.8***	0.4**	-0.27				
CON-5	0.8***	1.0***	-0.27				
CON-6	0.2**	0.6*	-1.00				
CON-7	0.4**	1.0	-0.45				
CON-8	0.6*	0.6*	-0.27				
CON-9	0.4**	0.6*	-0.64				
CON-10	0.2**	0.6*	-0.82				
CON-11	1.0***	0.6*	-0.09				

CON-12	0.8***	1.0***	-0.27
CON-13	0.4**	0.8***	-0.64
CON-14	0.4**	0.8***	-0.45
CON-15	0.6*	1.0***	-0.27
CON-16	0.0**	0.6*	-0.64
CON-17	0.6*	0.8	-0.45
CON-18	0.4**	0.4**	-0.45

Note: Values with asterisk * , **, and ***, are fair, good, and excellent, respectively.

From the expert reviewers' responses, probability of chance of agreement (P_C) and magnitude of strength of agreement, Fleiss κ [31, 32] values for Clarity and Relevance of each item are presented in Table 7. Over half of the κ values for Clarity, was considered *excellent*, while there was a quarter of the κ values for Relevance that were considered *excellent*. Overall, there were no items within the benchmark of *good*. For both κ values of Clarity and Relevance, there were ten items which fit within the benchmark to be considered *fair*. These values can indicate the reliability and consistency of experts' responses for each survey item. Paired with the previous analyses, Fleiss' κ gives further insight of initial validation of the instrument.

Table 4 – *Content Validity Evidence (Creating Value)*

Item	***I-CVI$_R$***	***I-CVI$_C$***	***CVR***	***S-CVI$_R$***	***S-CVI$_C$***	***AV-CVI$_R$***	***AV-CVI$_C$***
CV-1	0.0**	0.2**	-0.45	0.10	0.48	0.45	0.78
CV-2	0.3**	1.0***	-0.45				
CV-3	0.3**	1.0***	-0.45				
CV-4	0.3**	0.4**	-0.45				
CV-5	0.3**	1.0***	-0.45				
CV-6	0.5*	0.8***	-0.45				
CV-7	0.0**	1.0***	-0.64				
CV-8	0.3**	0.6**	-0.64				
CV-9	0.5*	1.0***	-0.64				
CV-10	0.5*	0.8***	-0.64				
CV-11	1.0***	1.0***	-0.09				
CV-12	0.5*	0.4**	-0.27				
CV-13	0.5*	0.8***	-0.45				
CV-14	1.0***	1.0***	-0.09				
CV-15	0.3**	1.0***	-0.45				
CV-16	0.3**	1.0***	-0.82				
CV-17	0.8***	0.6*	-0.45				
CV-18	0.8***	0.6*	-0.64				
CV-19	0.5*	1.0	-0.45				

CV-20	0.5*	0.6*	-0.27
CV-21	0.8***	0.6*	-0.27

Note: Values with asterisk * , **, and ***, are fair, good, and excellent, respectively.

Table 5 – *Probability of Chance of Agreement (P_C) & Fleiss'* κ *(Clarity and Relevance)*

	Item	***Pc-Clarity***	***Pc-Relevance***	***κ - Clarity***	***κ - Relevance***
Connection	CON-1	0.16	0.31	0.76**	0.42*
	CON-2	0.16	0.31	0.76**	0.13
	CON-3	0.03	0.03	-0.03	-0.03
	CON-4	0.31	0.16	0.13	0.76**
	CON-5	0.03	0.16	1.00**	0.76**
	CON-6	0.31	0.16	0.42*	0.05
	CON-7	0.03	0.31	1.00**	0.13
	CON-8	0.31	0.31	0.42*	0.42*
	CON-9	0.31	0.31	0.42*	0.13
	CON-10	0.31	0.16	0.42*	0.05
	CON-11	0.31	0.03	0.42*	1.00**
	CON-12	0.03	0.16	1.00**	0.76**
	CON-13	0.16	0.31	0.76**	0.13
	CON-14	0.16	0.31	0.76**	0.13
	CON-15	0.03	0.31	1.00**	0.42*
	CON-16	0.31	0.03	0.42*	-0.03
	CON-17	0.16	0.31	0.76**	0.42*
	CON-18	0.31	0.31	0.13	0.13
Creating Value	CV-1	0.16	0.16	0.05	0.05
	CV-2	0.03	0.31	1.00**	0.13
	CV-3	0.03	0.31	1.00**	0.13
	CV-4	0.31	0.31	0.13	0.13
	CV-5	0.03	0.31	1.00**	0.13
	CV-6	0.16	0.31	0.76**	0.42*
	CV-7	0.03	0.16	1.00**	0.05
	CV-8	0.31	0.31	0.42*	0.13
	CV-9	0.03	0.31	1.00**	0.42*
	CV-10	0.16	0.31	0.76**	0.42*
	CV-11	0.03	0.03	1.00**	1.00**
	CV-12	0.31	0.31	0.13	0.42*
	CV-13	0.16	0.31	0.76**	0.42*
	CV-14	0.03	0.03	1.00**	1.00**
	CV-15	0.03	0.31	1.00**	0.13
	CV-16	0.03	0.31	1.00**	0.13
	CV-17	0.31	0.16	0.42*	0.76**

	CV-18	0.31	0.16	0.42*	0.76**
	CV-19	0.03	0.31	1.00**	0.42*
	CV-20	0.31	0.31	0.42*	0.42*
	CV-21	0.31	0.16	0.42*	0.76**

Note: Values with asterisk * and **, are fair and excellent, respectively.

5. Discussion

Expert reviewers provided insight into both the face level appropriateness of items within the indirect assessments as well as quantitative scaling for each item's relevance, clarity, and essentiality. This data will be used to inform item reduction of the indirect assessments to improve the quality of the instruments. Thus, validation evidence for face validity indicated that three primary items should be removed from the indirect assessments with CON-10, CV-6, and CV-16 falling beneath the established benchmarks for item level inclusion. The metrics for content validation, which guided item reduction, are more complex and must be analyzed based on expert reviewers' focus on relevance, essentiality, and clarity.

Expert reviewers' evaluation of relevance indicated an overall low scoring at both the item and scale levels. Typically, items falling below the benchmarks would be removed. However, due to the lack of agreement established by the Kappa value, relevance alone is not an ideal criterion for item reduction. Several factors may have contributed to these overall low scores and poor agreement across experts. One potential mechanism being the diversity of "experts" available within KEEN that could pose a continuum of perceived relevance that would vary. More specifically, though there is a theoretical "definition" of constructs, "experience-based definition" may have influenced their the judgements for the indirect assessment's items. Additionally, the implementation of the 3Cs of EML contain creative and expansive varieties that cause "experts" to engage with these theoretical constructs in different capacities influencing their reviews of the item. Despite these variations, this research study remains confident in the relevance of each item, as they were directly developed from KEEN's definitions of connections and creating value, again warranting for relevancy to not be the most discriminative review criteria for the content validity analysis.

The second focus for this content validity analysis interrogated the clarity of the items within the indirect assessments. For the indirect assessment for Connections, CON-10 did not establish evidence of clarity in content validity affirming the recommendation to remove the item. However, eight additional items fell beneath the required benchmarks for I-CVI_C including CON-3, CON-4, CON-6, CON-8, CON-9, CON-10, CON-11, CON-16, and CON-18. Similarly, the indirect assessment for Creating Value identified eight more items, beyond CV-6 and CV-16, that demonstrated poor face validity. These items, which fell below the benchmarks for Clarity in Content Validity, include CV-1, CV-4, CV-8, CV-12, CV-17, CV-18, CV-20, and CV-21.

The third focus expert reviewers were asked to score was on the essentiality of each item in which the Content Validity Ratio (CVR) was the primary metric. From the results in Table 6, all CVR values for both indirect assessments fall beneath zero indicated a spectrum of weak to strong disagreement across expert reviews. This result is synergistic of the previously made

conclusion about relevance in that the disagreement could be anchored in the variety of experts within the KEEN Network. The CVR results do not prove to be as useful of a metric as originally hypothesized. Instead, with strong disagreement reifying in the results of the CVR, it becomes clear that the quantitative analysis used within this study is insufficient on its own to draw meaningful conclusions from the expert review's judgements. Instead, future item reduction will require the infusing of the open-ended qualitative comments reviewers left to supplement justification. We will use these findings to support future item reductions; however, the qualitative open-ended comments from the reviewers must be included to fully inform item reduction. More specifically, the results of this analysis utilizing purely quantitative analysis demonstrate insufficient evidence to support item reductions and requires further investigation into the qualitative data provided by expert reviewers. Despite literature support that five expert reviewers provide adequate means for face and content validity (McCoach et al., 2013), the expansive of how experts engage with EM remains limiting to knowing whether saturation occurred in data collection. As a result, continuous reviews will be collected to support ongoing improvement of the indirect assessments. Limitations for the study are included further in the discussion; however, revisiting the CVR with a larger pool of expert reviewers would be necessary to substantiate any claims regarding the essentiality of items within the indirect assessments.

Numerous limitations of the study existed. The first of these being the aforementioned spectrum of "experts" within KEEN. The KEEN network primarily focuses on the teaching and learning of EM and is not primarily a research focused collection of professionals. As a result, the ways one sees a component of EML as relevant or essential would be grounded more in their interpretation of their experiences with EM. However, the depth and spectrum of EM makes it challenging for all experts to share a common experience-based definition. A second major limitation of the study was in the survey design for collecting expert reviewer judgement. The Likert scaling being used for reviewers to score how relevant, clear, and essential items was not consistent. Relevancies include a 4-point Likert scale whereas the others only allowed a 3-point. Though this seems pedantic, the simple inclusion of more options makes it challenging to make claims across the three dimensions of the expert's judgement. This also could be a component for why Relevance scores underperformed given the elevated discrimination occurring. Despite these limitations of the research design, we remain confident that the evidence established through this investigation informs the progression of necessary assessment tools for Connections, Creating Value, and EML at large.

6. Conclusion

This paper presents the initial investigation of validation evidence for the indirect assessments measuring Connections & Creating Value. The intention of this study was to establish meaningful evidence to inform item reductions for the indirect assessments on the basis of face validity and content validity. The analysis was proven to be insufficient in establishing this evidence given the use of solely quantitative data and analyses. Future work will look to integrate a mixed methods approach by incorporating expert reviewer qualitative judgements to refine the evidence and better inform item reduction. The quality of this result is important in informing the reduction of items for these indirect assessments to elevate the instruments quality on the basis of experts reviews. This item reduction will heighten the overall quality of the

indirect assessments and will set the stage to begin exploring future validation constructs such as construct validity. The results of this study advance set of unique assessment tools capable of supporting engineering educators interested in adopting Entrepreneurial Minded Learning into their curriculum.

References

[1] Committee on Extraordinary Engineering Impacts on Society, National Academy of Engineering, and National Academies of Sciences, Engineering, and Medicine, *Impacts of National Science Foundation Engineering Research Support on Society*. Washington, D.C.: National Academies Press, 2024, p. 27873. doi: 10.17226/27873.

[2] C. B. Iturbe, L. Ochoa, M. C. J. Castello, and J. M. C. Pelayo, "Educating the engineer of 2020: Adapting engineering education to the new century," *IEEE Eng. Manag. Rev.*, vol. 37, pp. 11–11, 2009.

[3] A. R. Bielefeldt, "Professional Social Responsibility in Engineering," in *Social Responsibility*, I. Muenstermann, Ed., InTech, 2018. doi: 10.5772/intechopen.73785.

[4] D. A. MacKenzie and J. Wajcman, Eds., *The Social shaping of technology: how the refrigerator got its hum*. Milton Keynes ; Philadelphia: Open University Press, 1985.

[5] C. Mitcham, "Thinking through Technology: The Path between Engineering and Philosophy," *Bibliovault OAI Repos. Univ. Chic. Press*, vol. 36, Oct. 1995, doi: 10.2307/3106947.

[6] V. C. McGowan and P. Bell, "Engineering Education as the Development of Critical Sociotechnical Literacy," *Sci. Educ.*, vol. 29, no. 4, Art. no. 4, Aug. 2020, doi: 10.1007/s11191-020-00151-5.

[7] A. McAlister and S. Lilly, "Integrating Technical and Social Issues in Engineering Education: A Justice Oriented Mindset," Jun. 2023. doi: 10.18260/1-2–43783.

[8] B. Hynes and I. Richardson, "Entrepreneurship education: A mechanism for engaging and exchanging with the small business sector," *Educ. Train.*, vol. 49, pp. 732–744, Nov. 2007, doi: 10.1108/00400910710834120.

[9] "Engineering_Students'_Experien.pdf."

[10] E. Liventsova, T. Rumyantseva, and E. Syriamkina, "Development of Social and Entrepreneurial Skills of Students of Engineering and Technical Specialties in the Modern University," *MATEC Web Conf.*, vol. 79, p. 01018, 2016, doi: 10.1051/matecconf/20167901018.

[11] J. Hess and N. Fila, "The Development and Growth of Empathy Among Engineering Students," in *2016 ASEE Annual Conference & Exposition Proceedings*, New Orleans, Louisiana: ASEE Conferences, Jun. 2016, p. 26120. doi: 10.18260/p.26120.

[12] "What is KEEN? | Engineering Unleashed." Accessed: Feb. 20, 2025. [Online]. Available: https://engineeringunleashed.com/what-is-keen

[13] "Entrepreneurial Mindset | Engineering Unleashed." Accessed: Feb. 20, 2025. [Online]. Available: https://engineeringunleashed.com/mindset

[14] A. Huang-Saad, C. Bodnar, and A. Carberry, "Examining Current Practice in Engineering Entrepreneurship Education," *Entrep. Educ. Pedagogy*, vol. 3, no. 1, pp. 4–13, Jan. 2020, doi: 10.1177/2515127419890828.

[15] P. Shekhar and A. Huang-Saad, "Examining engineering students' participation in entrepreneurship education programs: implications for practice," *Int. J. STEM Educ.*, vol. 8, no. 1, p. 40, Dec. 2021, doi: 10.1186/s40594-021-00298-9.

[16] J. Qosaj, D. Corti, and S. Terzi, "Innovation & Entrepreneurship in Engineering Curricula: Evidences from an International Summer School," in *Advances in Production Management Systems. Production Management Systems for Responsible Manufacturing, Service, and Logistics Futures*, vol. 690, E. Alfnes, A. Romsdal, J. O. Strandhagen, G. Von Cieminski, and D. Romero, Eds., in IFIP Advances in Information and Communication Technology,

vol. 690. , Cham: Springer Nature Switzerland, 2023, pp. 461–475. doi: 10.1007/978-3-031-43666-6_32.
[17] M. T. Kane, “Validating the Interpretations and Uses of Test Scores,” *J. Educ. Meas.*, vol. 50, no. 1, pp. 1–73, Mar. 2013, doi: 10.1111/jedm.12000.
[18] S. Messick, “Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning.,” *Am. Psychol.*, vol. 50, no. 9, Art. no. 9, 1995, doi: 10.1037/0003-066X.50.9.741.
[19] S. Messick, “Meaning and Values in Test Validation: The Science and Ethics of Assessment,” *Educ. Res.*, vol. 18, no. 2, Art. no. 2, 1989, doi: 10.2307/1175249.
[20] M. Peeters and B. Martin, “Validation of Learning Assessments: A Primer,” *Curr. Pharm. Teach. Learn.*, vol. 9, pp. 925–933, Sep. 2017, doi: 10.1016/j.cptl.2017.06.001.
[21] W. M. Lim, “A typology of validity: content, face, convergent, discriminant, nomological and predictive validity,” *J. Trade Sci.*, vol. 12, no. 3, Art. no. 3, Jan. 2024, doi: 10.1108/JTS-03-2024-0016.
[22] M. G. K. Dijksterhuis, I. Jozwiak, D. D. M. Braat, and F. Scheele, “An exploratory study into the effect of time-restricted internet access on face-validity, construct validity and reliability of postgraduate knowledge progress testing.,” *BMC Med. Educ.*, vol. 13, p. 147, Nov. 2013, doi: 10.1186/1472-6920-13-147.
[23] E. S. Alias, M. Mukhtar, and R. Jenal, “Instrument Development for Measuring the Acceptance of UC&C: A Content Validity Study,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 4, Art. no. 4, 2019, doi: 10.14569/IJACSA.2019.0100422.
[24] D. B. McCoach, R. K. Gable, and J. P. Madura, *Instrument Development in the Affective Domain: School and Corporate Applications*. New York, NY: Springer New York, 2013. doi: 10.1007/978-1-4614-7135-6.
[25] M. M. de Lyra, S. Youssef, and K. M. Kecskemety, “Work-In-Progress: Validation of Connections and Creating Value Assessments”.
[26] R. Likert, “A technique for the measurement of attitudes.,” *Arch. Psychol.*, vol. 22 140, pp. 55–55, 1932.
[27] Department of Medical Education, School of Medical Sciences, Universiti Sains Malaysia, Kelantan, MALAYSIA and M. S. B. Yusoff, “ABC of Response Process Validation and Face Validity Index Calculation,” *Educ. Med. J.*, vol. 11, no. 3, Art. no. 3, Oct. 2019, doi: 10.21315/eimj2019.11.3.6.
[28] Department of Medical Education, School of Medical Sciences, Universiti Sains Malaysia, MALAYSIA and M. S. B. Yusoff, “ABC of Content Validation and Content Validity Index Calculation,” *Educ. Med. J.*, vol. 11, no. 2, Art. no. 2, Jun. 2019, doi: 10.21315/eimj2019.11.2.6.
[29] V. Zamanzadeh, A. Ghahramanian, M. Rassouli, A. Abbaszadeh, H. Alavi-Majd, and A.-R. Nikanfar, “Design and Implementation Content Validity Study: Development of an instrument for measuring Patient-Centered Communication,” *J Caring Sci*, vol. 4, no. 2, Art. no. 2, 2015, doi: 10.15171/jcs.2015.017.
[30] C. H. LAWSHE, “A QUANTITATIVE APPROACH TO CONTENT VALIDITY,” *Pers. Psychol.*, vol. 28, no. 4, Art. no. 4, 1975, doi: https://doi.org/10.1111/j.1744-6570.1975.tb01393.x.
[31] W. P. F. Robert Wilson and D. A. Schumsky, “Recalculation of the Critical Values for Lawshe’s Content Validity Ratio,” *Meas. Eval. Couns. Dev.*, vol. 45, no. 3, Art. no. 3, 2012, doi: 10.1177/0748175612440286.

[32] N. Samad, M. Asri, M. A. Mohd Noor, M. Mansor, and T. Jumahat, “Measuring the Content Validity of Middle Leadership Competence Model using Content Validity Ratio (CVR) Analysis,” *Int. J. Bus. Technol. Manag.*, vol. 5, pp. 134–144, Dec. 2023, doi: 10.55057/ijbtm.2023.5.4.12.
[33] J. L. Fleiss, “Measuring nominal scale agreement among many raters.,” *Psychol. Bull.*, vol. 76, no. 5, Art. no. 5, Nov. 1971, doi: 10.1037/h0031619.
[34] J. L. Fleiss, J. Cohen, and B. S. Everitt, “Large sample standard errors of kappa and weighted kappa.,” *Psychol. Bull.*, vol. 72, no. 5, Art. no. 5, Nov. 1969, doi: 10.1037/h0028106.