Faculty Online Teaching Effectiveness Scale (FOTES): Instrument Development and Content Validation

Yea-Jyh Chen[1], PhD RN, Yvonne M. Smith[2], PhD APRN-CNS
Andrea Warner Stidham[3], PhD RN

[1] College of Health and Human Services School of Nursing, University of North Carolina at Wilmington, Wilmington, NC

[2] Department of Nursing, Baldwin Wallace University, Berea, OH

[3] College of Nursing, Kent State University, Kent, OH

Correspondence. Yea-Jyh Chen, PhD, RN, College of Health and Human Services School of Nursing, University of North Carolina at Wilmington, 601 S College Road, Wilmington, NC 28403 (Phone: 910-962-0584; Fax: 910-962-4921; chenyj@uncw.edu).

Email/ORCID. Yea-Jyh Chen chenyj@uncw.edu https://orcid.org/0000-0003-1946-0683; Yvonne Smith ysmith@bw.edu https://orcid.org/0000-0001-5130-4870; Andrea Warner Stidham awarner@kent.edu https://orcid.org/0000-0003-1783-9980

Abstract

This study aimed to develop and psychometrically test the Faculty Online Teaching Effectiveness Scale (FOTES) based upon both student and faculty perspectives of online teaching and learning in higher education. Online teaching effectiveness is a crucial component of quality education, but it has not been well-defined conceptually, and few studies have been conducted, using relevant domains, to accurately measure online teaching effectiveness. The impact of online course delivery on teaching effectiveness remains unclear. An exploratory sequential mixed methods design was employed with three phases of instrument development and psychometric testing. The FOTES comprises 50 items in seven domains: teaching philosophy, self-efficacy, relationships, course content, learning activities, teaching practices, and satisfaction. The instrument underwent initial testing, yielding positive expert appraisals with good-excellent psychometrics. All domains of the scale were significantly correlated, except for teaching philosophy. The preliminary results of the FOTES provide the empirical evidence to advance additional psychometric validation. This newly developed instrument has the potential to enhance faculty capacity and skill in self-evaluating their teaching effectiveness in online courses, providing a valid and reliable measure. The resulting instrument is poised to promote outcome evaluation and strengthen teaching and learning processes.

Keywords. Effective teaching, online teaching effectiveness, content validity, psychometric validation, instrument development, self-evaluation of online teaching effectiveness

Faculty Online Teaching Effectiveness Scale (FOTES): Instrument Development and Content Validation

1. Introduction

Online education has evolved over time but was propelled into contemporary K-12 and higher education worldwide due to the COVID-19 pandemic crisis. The past few decades have brought a noticeable transition from traditional classroom education to using a virtual platform to fit the diverse needs of students. In the U.S., college distance education experienced a 29% increase between 2012 and 2018 (Ruiz & Sun, 2021). Enrollments in online courses surged during the pandemic, increasing 93% between fall 2019 and fall 2020 (Lederman, 2021). This rapid growth resulted in approximately 9.4 million students taking online or distance learning courses in 2021, representing 61% of the total undergraduate enrollment (National Center for Education Statistics, 2023).

An online course is operationally defined as a course in which all learning activities and content delivery occur completely online, synchronously or asynchronously, without in-person class sessions or on-campus requirements. Similarly, online learning, evolving into the mainstream of contemporary higher education, adds a component of technology-based learning in the definition (Crews et al., 2015; Singh & Thurman, 2019). In today's ever-changing world, teaching and learning complexity continues to drive classroom learning to the online environment beyond the traditional geographic boundaries and time constraints. Integrated into nursing education, online teaching has engendered significant challenges to faculty in their teaching (Islam et al., 2015; Philipsen et al., 2019; Smith & Crowe, 2017). For instance, compared to face-to-face classroom teaching, unique challenges include communication, absence of classroom dynamics, understanding of online pedagogy, use of media and technology, and

formative assessment of student learning. These factors directly impact teaching effectiveness (TE), a widely accepted indicator of course quality that is neither defined nor well-studied in the context of online teaching and learning.

TE is a continuing concern in online education despite previously identified challenges. Effectiveness in teaching has been addressed primarily in the traditional K-12 face-to-face classroom (Doan et al., 2019; Lauermann & ten Hagen, 2021). With less attention focused on defining TE in higher education, the school system and accrediting body evaluation processes' have not kept pace with the growth of online education. Additionally, faculty self-evaluation of their own teaching is an important component of teacher development and essential to being an effective teacher (Hattie, 2015). Teacher's understanding of their impact on higher education results in enhanced faculty teaching and maximized student learning (Hattie, 2015). Thus, it is important to consider faculty teaching self-evaluation and ensure that teaching quality in online courses equates to the quality of face-to-face classroom education (Elumalai et al., 2020; Kang, 2012).

1.1 Definition

Historically, student achievement gains such as test scores, academic grades, and passing rates have been used as a primary indicator of teaching effectiveness (TE). The need to use other methods of evaluating TE has been identified in the literature (Berk, 2013a; Esarey & Valdes, 2020; Makkonen, 2016; Smith et al., 2021; Uttl et al., 2017) suggesting we "opt out" of the practice of using primarily student outcomes. Similar terminologies of "teacher effectiveness," "educational effectiveness," "teacher quality," and "good teaching" appear in the literature. These are difficult to distinguish from TE, but their focus is mainly on student outcomes, characteristics of an individual teacher in the classroom, and instructional and classroom-

oriented programs or schools (Cinches et al., 2017; Doan et al., 2019). Layne (2012) and Berk (2013a) challenged the idea that a focus on teaching detracts from a focus on learning and advocated for a focus on both to fortify the cohesiveness of teaching and learning. A conceptual definition of TE and a shift in focus from face-to-face to online education are lacking in recent literature. Some educators offer best practices for online teaching (Akram et al., 2021; Berk, 2013b; Price et al., 2016), providing insight into how quality online teaching is perceived. Frazer et al. (2017) identified three themes —TE, indicators of quality, and student success — that are intricately linked in the online environment. However, their findings are limited to Ph.D.-prepared faculty from one university offering only asynchronous online courses, with an unclear conceptual definition of TE. According to Smith and Crowe (2017), nursing faculty perceived effective online teaching as being inherently linked to forming relationships with students and having a strong presence in the online classroom. Faculty identified the establishment of these relationships as their greatest challenge. These qualitative studies provide a foundation for further exploration of TE in the online environment.

1.2 Current instrumental measurements

Current instruments that attempt to measure components of TE include the Community of Inquiry (COI) Survey, the Nursing Clinical Teaching Effectiveness Inventory (NCTEI), the Student Evaluation of Online Teaching Effectiveness (SEOTE) and the Online Teaching Effectiveness Scale (OTES). The COI Survey measures the degree to which a community of inquiry has been created in the online environment (Arbaugh et al., 2008). The NCTEI is specific to TE in the clinical environment and is not transferable to the online learning environment (Morgan & Knox, 1987). Developed from Chickering and Gamson's (1987) Seven Principles of Effective Teaching, the SEOTE is used to gather student feedback on online teaching practices

(Bangert, 2008). The four-factor model of the OTES was evidence-based and designed for students, rather than instructors, to evaluate faculty OTE (Reyes-Fournier et al., 2020). However, this instrument excludes the potential student-associated effects on the evaluation response. Overall, each of these instruments evaluate components of online teaching from the student perspective and have shown a positive contribution to effective online courses. Despite these efforts, the links between these components and the outcome of instructors' perceived TE in online education remains unclear.

1.3 Gaps in literature and practice

The interpretation of online classroom observation and teacher presence are different from the traditional aspects and definition, which also constrains the application and evaluation of online TE. Moreover, existing instruments are restricted to a single dimension, such as using students' grade achievement, measuring student satisfaction, or relying solely on student evaluations of TE (Arbaugh et al., 2008; Bangert, 2008; Chianese, 2015; Esarey & Valdes, 2020). Academia lacks a science-based instrument for faculty use in fully measuring online TE. Expanding our understanding of quality teaching by developing an instrument to measure online TE, that encompasses faculty's perspective, is essential to minimize inconsistent teaching and maximize learning outcomes.

2. Study aim and research questions

Derived from the previous thematic data results (Smith et al., 2021), this study aimed to develop and initially test psychometrics of the new instrument of the Faculty Online Teaching Effectiveness Scale (FOTES). There were three research questions tested:

1. To what extent were the relevant and conceptual-based items generated associated with online TE?
2. What was the content validity of the developed instrument?
3. How well did the instrument measure all aspects of the domains of faculty online TE?

3. Materials and methods

3.1 Methodology

3.1.1 Design

An exploratory sequential mixed methods design (Creswell, 2015) was conducted to methodologically develop and initially test an instrument to measure online TE. The study design comprised three phases: a two-stage process to assess content validity (Lynn,1986), and pilot testing the instrument. Phase I instrument construction was conducted to develop the individual items to form the item pool of the scale and identify relevant domains of faculty OTE. During phase II content evaluation, the scale items were judged and quantified by a panel of experts to determine content validity (Grant & Davis, 1997). The initial psychometric properties of the scale were determined, using a small sample of university faculty, for pre-test in phase III. Figure 1 summarizes three study phases, procedures, and results.

3.1.2 Instrument development and item generation

After conducting the qualitative study with focus group interviews, the pragmatic characteristics derived from online teaching and learning experiences and perspectives were precisely defined in relation to OTE (Smith et al., 2021). The qualitative results along with the synthesis of TE literature were used to inductively generate an item pool aligned with the operational domains. Throughout the seven months undergoing this process, the study

researchers carefully transformed thematic codes and quotes into survey items and repeatedly inspected each item to accurately reflect the pivotal meaning of the data. To fully contextualize Faculty perceptions of TE, we adopted a unique approach by comparing faculty perception of TE with the perceptions of students taking online courses during item generation. The initial questionnaire contained 63 items with a 6-point Likert scale evaluating from 1 (strongly disagree) to 6 (strongly agree), plus a possible option of "I do not participate". The eight domains addressed by the initial scale included relationships, motivating students, teaching philosophy, self- efficacy, course content, learning activities, teaching practices, and satisfaction.

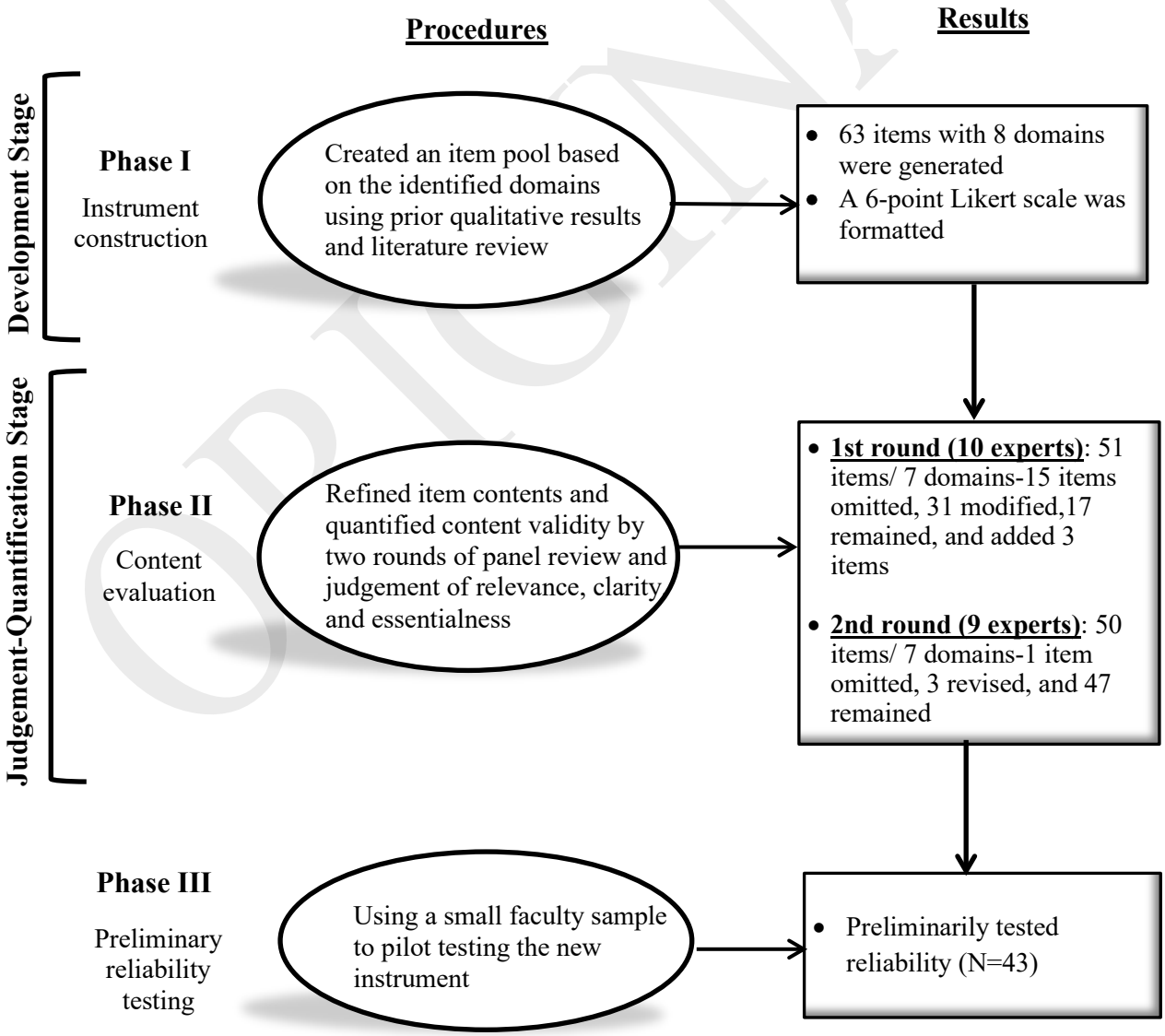

Fig. 1. Summary of study three phases of content validation process and results using a two-stage process for content determination (Lynn, 1986) following a pilot test.

3.1.3 Content validation and refinement employing an expert panel

During phase II, the judgement-quantification stage (Lynn, 1986), a panel of 10 experts from different disciplines (5 nursing and 5 non-nursing) and universities were identified and invited to participate. To efficiently measure content validity, this panel consisted of six online education content experts and four instrumentation experts. The experts independently reviewed the scale items, item-to-domain content, and the entire set of the questionnaire response scales, completing their reviews via the Qualtrics platform. The criteria for review were provided as a guide to assure the item review and scoring were approached in a consistent manner (Davis, 1992; Polit et al., 2007). The elements of the assessment criteria included item relevance, clarity, and essentialness with 1-4 points for relevance and clarity, and 1-3 points for essentialness. The expert responses also included specific comments with rationales, including suggestions in item ordering, wording of the questions, response choices and styles, and scale instruction (Davis, 1992; DeVellis, 2017).

The first expert review by ten panelists required a minimum CVR of 0.62 and 0.78 for the second review with nine experts to satisfy a 5% level of significance (Lawshe, 1975). As such, greater than six out of ten expert members needed to agree on the essentialness of each item. A second-round panel review was conducted for the identified eight-item subset of the scale (Davis, 1992; Polit et al., 2007). The importance of each item was also considered in relation to the construct and the rating of the expert as “useful but not essential”. An individual expert approach was further employed to confirm the amended items and the questionnaire scale, as a whole, to achieve consensus and ensure the integrity of the concept defined. This additional assurance was obtained to ensure study rigor and quality and validation of the content measure, from individual experts, before finalizing the scale (Grant & Davis, 1997; Polit et al., 2007). The

substantial improvement on the scale items was acknowledged after the subsequent post-hoc expert review.

Following each review round, the item consensus was first assessed, and the low item agreement (< 0.80) was re-evaluated with either revising or removing the items to meet the acceptable agreement (Davis, 1992). The modified Kappa statistics (K*) was applied to adjust for inter-rater agreements for the expert panel (Polit et al., 2007; Wynd et al., 2003). Content validity was computed to determine how well each item represented the construct of online TE, including item content validity index (I-CVI) and scale-level CVI (S-CVI; Davis, 1992) and content validity ratio (CVR; Lawshe, 1975). Excellent content validity is considered by achieving a cut-off point of I-CVI $\geq$ 0.78 and S-CVI $\geq$ 0.80-0.90 (Polit et al., 2007). Items with I-CVI < 0.70 and/or CVR< 0.62/0.78 were eliminated and other items between I-CVI 0.70- 0.79 were modified or removed. Revisions were made accordingly, based upon the experts' response results and feedback, to align with the construct of faculty online TE. The validity measures and inter-rater agreement were examined for two rounds of expert reviews.

3.1.4 Pilot testing the initial instrument

At phase III, a small faculty group was recruited, using a modified Dillman approach, for pretesting the newly developed instrument. Nursing faculty from all public universities in a Midwest state (identified through university webpages) were invited via email. Faculty who were interested in the study consented and completed a demographic survey, using Qualtrics Survey Software, to determine their study eligibility. Eligible participants with either full-time or part-time faculty position and online teaching experience completed the instrument questionnaire and responded to several cognitive-oriented questions, such as understanding of each question, error item information of measured constructs, or improper description of questions (Chaudhary

& Israel, 2014; DeVellis, 2017). Question response patterns, length, and flow were also monitored and described. This pilot testing result was used to determine consistency over item responses and whether further modifications to an item or the instrument was needed.

3.2 Ethical considerations

University Institutional Review Board (IRB No. 17-524) approval was obtained before approaching potential study participants. The study informed consents and survey data were stored electronically with password protection and encryption on the University server. Confidentiality was maintained by using study ID codes for participants and data were reported in the aggregate form. Only study researchers and team members had access to data files.

3.3 Data analyses

The electronic data were stored in Qualtrics and exported into the Statistical Package of the Social Sciences (SPSS; version 24.0, IBM Corp.) for analysis. Participants' characteristics were reported using mean, standard deviation, and range. Item inter-rater agreements were determined by percentages for each panel review. The item-level content validity index (I-CVIs) and scale-level content validity index/averaging (S-CVI) were calculated by the formulas to validate item content. According to the criteria (Polit et al., 2007), the I-CVI was based on the number of experts who rated the item as relevant and clear (%); and the S-CVI was computed by translating the 4-point score to dichotomous kappa statistics for proportional agreement among experts, e.g., relevant vs. not relevant. Modified Kappa statistics (K^*) were calculated using the formula, $(1\text{-CVI} - P_c)/ (1-P_c)$, P_c= Probability of a chance occurrence (Polit et al., 2007). The calculated probability was then interpreted using the standard ranges of Kappa statistic by Cicchetti (1984) and Fleiss (1981). The study data in phase III resulted from pilot testing the scale. Item analyses by screening the item-scale and scale-total correlation matrices was used to

determine acceptable internal reliability of the instrument. Due to non-normal distribution detected in most item-domain data, *Spearman's* rank-order correlation (*rho*) was applied in understanding the inter-item and domain-total scale correlations. Cronbach's *alpha* was calculated to indicate consistency on the measurement (Nunnally & Berstein, 1994). The study statistical significance was set at 0.05.

4. Results

4.1 Development of the conceptual-based instrument

After the first panel review of the 63 initial survey items, 22 items (35%) remained unchanged or with slight modification, 29 items were revised, and 12 items were discarded. The revised scale was comprised of 51 items and 7 domains. One domain of 'motivating students' was eliminated from the original version of the scale. At the post-hoc review, 47 of 51 total items (92%) remained as presented, or with minor rewording to enhance context specificity. Only three items needed major revision and one item was excluded. This eliminated item, "I am comfortable with my skill level with the associated technologies" was determined less essential to the FOTE based on the experts' comments and item essentialness below 0.78 (CVR= 0.33; see Table 1). The final version of the FOTES comprises 50 core items, in 7 domains, in accordance with the second post-hoc review presented in Table 2. A 6-point Likert scale was included to purposely evaluate elements of teaching effectiveness. The scale responses for how individual faculty perceived their online teaching range from 1 (very untrue) to 6 (very true) in three domains—teaching philosophy, self-efficacy, and satisfaction. In the remaining four domains, which relate to specific actions taken in teaching, the scale responses range from 1 (never) to 6 (very frequently). The higher the average score of all items, the greater online TE.

Table 1

Content evaluation of the final instrument after two-rounds of panel reviews.

Conceptual Domain	Item	Essentialness		Interpretation	Clarity			Relevance			Interpretation
		N_e	CVR		$N_{3,4}$	I-CVI	K*	$N_{3,4}$	I-CVI	K*	
D1: Teaching Philosophy	1	9	0.80	Remained	9	0.9	0.90	10	1.0	1.00	Excellent
	2	9	0.80	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	3**	8	0.78	Remained	8	0.89	0.89	9	1.0	1.00	Excellent
	4**	8	0.78	Remained	8	0.89	0.89	8	0.89	0.89	Excellent
D2: Self-efficacy	1	9	0.80	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	2**	9	1.00	Remained	9	1.0	1.00	9	1.0	1.00	Excellent
	3**	6	0.33	Excluded							
	4	8	0.60	Remained	9	0.9	0.90	9	0.9	0.90	Excellent
	5	9	0.80	Remained	9	0.9	0.90	10	1.0	1.00	Excellent
	6**	8	0.78	Remained	7	0.78	0.76	9	1.0	1.00	Excellent
	7**	8	0.78	Remained	9	1.0	1.00	9	1.0	1.00	Excellent
D3: Relationships	1	8	0.60	Remained	9	0.9	0.90	9	0.9	0.90	Excellent
	2	9	0.80	Remained	9	0.9	0.90	10	1.0	1.00	Excellent
	3	9	0.80	Remained	8	0.8	0.79	10	1.0	1.00	Excellent
	4	9	0.80	Remained	9	0.9	0.90	10	1.0	1.00	Excellent
	5	9	0.80	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	6	9	0.80	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	7	8	0.60	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
D4: Course Content	1	10	1.00	Remained	9	0.9	0.90	10	1.0	1.00	Excellent
	2	8	0.60	Remained	9	0.9	0.90	10	1.0	1.00	Excellent
	3	10	1.00	Remained	8	0.8	0.79	9	0.9	0.90	Excellent
	4	9	0.80	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	5	10	1.00	Remained	8	0.8	0.79	10	1.0	1.00	Excellent
	6	9	0.80	Remained	8	0.8	0.79	9	0.9	0.90	Excellent
D5: Learning Activities	1	10	1.00	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	2	9	0.80	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	3	8	0.60	Remained	9	0.9	0.90	9	0.9	0.90	Excellent
	4	9	0.80	Remained	9	0.9	0.90	10	1.0	1.00	Excellent
	5	9	0.80	Remained	9	0.9	0.90	10	1.0	1.00	Excellent
	6**	8	0.78	Remained	9	1.0	1.00	9	1.0	1.00	Excellent
	7	8	0.60	Remained	7	0.7	0.66	10	1.0	1.00	Good-Excellent
	8	9	0.80	Remained	9	0.9	0.90	9	0.9	0.90	Excellent
	9	10	1.00	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
D6: Teaching Practices	1	9	0.80	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	2	10	1.00	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	3	10	1.00	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	4	9	0.80	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	5	10	1.00	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	6	10	1.00	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	7	10	1.00	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	8	9	0.80	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	9	10	1.00	Remained	9	0.9	1.00	10	1.0	1.00	Excellent
	10**	9	1.00	Remained	9	1.0	1.00	9	1.0	1.00	Excellent
D7: Satisfaction	1	10	1.00	Remained	7	0.7	0.66	10	1.0	1.00	Good-Excellent
	2	9	0.80	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	3	8	0.60	Remained	9	0.9	0.90	10	1.0	1.00	Excellent
	4	9	0.80	Remained	9	0.9	0.90	10	1.0	1.00	Excellent
	5	8	0.60	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	6	10	1.00	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	7	8	0.60	Remained	8	0.8	0.79	9	0.9	0.90	Excellent
	8	10	1.00	Remained	10	1.0	1.00	10	1.0	1.00	Excellent
	50 Items			**S-CVI/Avg**		0.93			0.98		

Note: N_e=Number of experts rated the item as essential; CVR= Content validity ratio; $N_{3,4}$= Number of experts rated 3 or 4 (relevant or clear) in the item; I-CVI= Item-level content validity index; P_c= Probability of a chance occurrence; Modified Kappa statistic [K*]= Kappa designating agreement on relevance using the formula= (1-CVI – Pc)/ (1-Pc); S-CVI/Avg= Scale-level content validity index average calculation. **Eight question items were specifically re-evaluated by 9 experts at the second round of the panel review. The rest of the items used the score points computed from the first round of the expert panel (N=10).

4.2 Content validity of the developed instrument

The final FOTES obtained expert consensus, yielding positive appraisals on the retained content. Table 2 summarizes the item's essentialness in the CVR, which was increased to 98% from 70% (satisfactory range of 0.60-1.00; Lawshe, 1975) after the second expert review. The content was relevant, supported by excellent I-CVI (1.00; 0.89-1.00) and S-CVI/ Avg (0.98) which aligned with excellent K*(0.89-1.00) of multi-rater agreement. Content clarity was also validated by I-CVI (0.94; 0.70-1.00) and S-CVI/ Avg (0.93).

Table 2
Content validity and inter-rater agreement by two rounds of expert reviews.

Validity Evaluation		Standard Criteria	Expert Review	
			First Round	**Second Round**
Relevance	I-CVI (%) (Range)	> = 0.78	57/63 items (90.5%) 0.40 - 1.00	51/51 items (100%) 0.89 - 1.00
	S-CVI/Avg	> 0.9	0.92	0.98
	K* (Range)	> = 0.60	0.25 - 1.00	0.89 - 1.00
Clarity	I-CVI (%) (Range)	> = 0.78	53/63 items (84%) 0.50-1.00	48/51 items (94%) 0.70-1.00
	S-CVI/Avg	> 0.9	0.89	0.93
Essentialness	CVR (%) (Range)	> = 0.62	44/63 items (70%) -0.60 - 1.00	50/51 items (98%) 0.60 - 1.00

Note: I-CVI= Item-level content validity index using the criteria by Polit et al. (2007): >= 0.78 excellent, between < 0.78 and >=0.70 considering for revision, and <0.70 for deletion; S-CVI/Avg= Scale-level content validity index/Averaging calculation method. S-CVI/Avg must equal to 90% or higher (Polit et al., 2007); Modified Kappa statistics [K*] = Kappa designating agreement index on relevance. The interpretation used the *kappa* magnitude criteria by Cicchetti (1984) and Fleiss (1981): Poor= k < 0.40, Fair= k of 0.40-0.59, Good= k of 0.60-0.74, Excellent= k of 0.75-1.00; CVR= Content validity ratio using the guideline in Lawshe (1975): a minimum CVR of 0.62 for the review by 10 panelists.

4.3 Pilot testing

4.3.1 Sample participants

The initial response rate was 11.2% (72 out of 645 emails sent). Of these responses, 43 eligible online nursing faculty (60%) from eight state universities completed the survey and provided the available data for study analysis. Faculty participants were more likely to be mid-aged (82.4% of the faculty were less than 65 years old; average age of 53.9 $\pm$ *SD* 10.8), females (93%), Caucasian (83.7%), non-tenured (51.2%), associate/assistant professors (58.2%) with doctorate degree (65.1%). Most of them (*N*: 30, 69.8%) were senior faculty with 10 years or more teaching experience (*range*: 3-47 years) and the average of 8.9 years (*SD*: 5.4) engaged in teaching online courses. Over two thirds of the study faculty (67.4%) identified that they had used a tool to previously evaluate their teaching effectiveness, listing student evaluation surveys or university course evaluations for this purpose. Table 3 provides the summary of the participants' demographics.

4.3.2 Preliminary results

In phase 3, the study participants rated a high level of TE in their online course teaching (see Table 4). Of this study sample, a negatively skewed pattern was more likely to be seen in each domain score (Domain response: *Mdn*= 5.33-5.86 & *IQR*= 0.43-0.87), whereas the score of the entire FOTES scale was normally distributed (*M*= 5.50 $\pm$ *SD* 0.34). The measures between each domain and the FOTES were congruent and significantly correlated (inter-scale *rho*= 0.31-0.69, $p < 0.05$; subscale-to-total scale *rho*= 0.53-0.83, $p < 0.001$), except the domain of teaching philosophy. Teaching philosophy did not significantly correlate with the two domains of relationships and satisfaction ($p > 0.05$; Table 4). Good internal consistency indicated the scale

measured the same construct of faculty online TE; and 89% of the variance in the scale test was reliable (Cronbach's α= 0.89).

Table 3
Phase III demographics of faculty participants (N=43).

Demographic Characteristics	***N* (%)**
Age in years old ($M \pm SD$; N= 34)	53.9 ± 10.8
Range	32-70
Gender: Females	40 (93)
Ethnicity: Caucasian	36 (83.7)
Employing university	
University 1	18 (41.9)
University 2	12 (27.9)
University 3-8	13 (30.2)
Academic rank	
Professor & Associate professor	15 (34.9)
Assistant professor	14 (32.6)
Senior lecturer/ Associate lecturer/Lecturer	10 (23.3)
Other (e.g., instructor etc.)	4 (9.3)
Academic classification	
Tenured/ Tenure-track	17 (39.6)
Non-tenure track	22 (51.2)
Part time/Adjunct faculty	4 (9.3)
Highest education degree	
Master	15 (34.9)
Doctorate	28 (65.1)
Total years of teaching experience ($M \pm SD$; *Range*)	15.9 ± 10.3 (3-47)
Total years of online teaching experience ($M \pm SD$; *Range*)	8.9 ± 5.4 (1-25)
Previously used a tool to evaluate teaching effectiveness: Yes	29 (67.4)
Student evaluation surveys	16 (55)
Course evaluation by the university	6 (21)
Quality Matters	2 (7)
Peer evaluation	2 (7)
Others	3 (10)

Note: M, mean; SD, standard deviation.

Table 4
The subscale scores, item-scale correlations, and internal consistency of the FOTES.

Domain[a]	Subscale Score		Subscale-to-Subscale Correlation[b,c] (*rho* by range)	Subscale-to-Total Scale Correlation (*rho*)
	M* ± *SD* / *Mdn	**Item Response (range/IQR)**		
Domain 1	5.54 ± 0.47/ 5.67	4.25 – 6.00/ 0.75	0.35 – 0.55*	0.53***
Domain 2	5.32 ± 0.55/ 5.33	4.00 – 6.00/ 0.67	0.31 – 0.43*	0.62***
Domain 3	5.72 ± 0.32/ 5.86	4.71 – 6.00/ 0.43	0.43 – 0.69**	0.81***
Domain 4	5.73 ± 0.26/ 5.83	5.00 – 6.00/ 0.50	0.31 – 0.68*	0.69***
Domain 5	5.54 ± 0.47/ 5.67	4.11– 6.00/ 0.56	0.35 – 0.68*	0.72***
Domain 6	5.48 ± 0.46/ 5.60	4.20 – 6.00/ 0.60	0.32 – 0.69*	0.83***
Domain 7	5.24 ± 0.71/ 5.50	3.25 – 6.00/ 0.87	0.34 – 0.61*	0.76***
Total Scale		***M/ SD/ Mdn***	***Cronbach's Alpha***	
		5.50/ 0.34/ 5.54	0.89	

Note: FOTES: Faculty Online Teaching Effectiveness Scale; *Mdn*: Median; *M*: Mean; *SD*: Standard Deviation; *rho*: Spearman's correlation coefficient. [a]Domain 1: Teaching Philosophy, Domain 2: Self-efficacy, Domain 3: Relationships, Domain 4: Course Content, Domain 5: Learning Activities, Domain 6: Teaching Practices, Domain 7: Satisfaction. [b]All domains were correlated with other domains, except Domain 1. [c]Domain 1 was not significantly correlated with Domain 3 and Domain 7 (*rho*= 0.25 & 0.26, $p > .05$, respectively). * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$

5. Discussion

This study advances the science of education by developing the self-assessment FOTES instrument which involves the perspectives of both faculty and student with experience in online education. The results provide preliminary data of using the new instrument to evaluate faculty TE with evidence-based elements in online education. Solid expert validity and good psychometrics were demonstrated by a group of content and methodological experts and nursing

faculty participants. Our work addresses a substantial gap in scientific knowledge as no other instruments for faculty self-evaluation are available for measuring online TE in higher education, in general, or other healthcare-related fields, specifically. It also highlights insights the faculty perceived regarding their teaching performance in online courses and reinforces self-awareness of their online TE. Further, the availability of this instrument provides an alternative to the less accurate, extraneous evaluations currently used in many higher education settings.

Competently measuring faculty online TE is challenging. TE in online course learning is not a stable attribute, but instead fluctuates often by course design, instructor, and individual student. Thus, a reliable measure is essential in yielding a true estimate of the FOTE. The first steps of creating an instrument and content validation were accomplished following the guideline procedures of standardized measures (DeVellis, 2017; Grant & Davis, 1997; Lynn, 1986). Using a multi-dimensional approach, our developed instrument has explicitly demonstrated the content validity of using conceptual-based model from our qualitative evidence (Smith et al., 2021) to measure online TE for faculty teaching in higher education. The scale items were categorized by the refined construct initiatives of teaching philosophy, self-efficacy, relationships, course content, learning activities, teaching practices, and satisfaction. All the content domains were positively associated with each other in alignment with the experts' consensuses on the practical context or attribute being measured.

These domains are similar to the key dimensions reported to assess OTE (Akram et al., 2021; Reyes-Fournier et al., 2020). The four-factor domain items in the Online Teaching Effectiveness Scale (OTES), as developed by Reyes-Fournier et al. (2020), pertain to assessing instructors' presence, expertise, facilitation, and engagement, based upon students' responses. Akram and colleagues (2021) investigated a hypothetical, non-instrument oriented OTE model,

focusing on five specific factors and their relationships with OTE, including students' active and passive class participation, teacher skills and strategies, teacher training, teaching domain, and teaching perceptions. Recognized by the university faculty, these five key factors significantly explained 56% of the variances in teaching effectiveness by the OTE model. These two recent research results have established the scientific evidence and congruence, in part, with the content domains generated in our study. In comparison, our work of developing the FOTES, with the defined domains and items, is more inclusive and beneficial to understanding faculty's individual reflection and evaluation of their own teaching practice. This self-report instrument is a composite of multiple OTE dimensions, tangible conceptual themes from our previous work (Smith et al., 2021), and evidence-based reviews, constructed to measure faculty perceived OTE rather than a more typical student-oriented perceptions.

We noticed that our preliminary data showed no association of two domains of relationships (domain 3) and satisfaction (domain 7) with teaching philosophy (domain 1; $p > 0.05$). The result can be explained with the inadequate item fit or the domain variable in relation to faculty OTE. This could imply that an item or items in the domain set may be incongruent with the measures of the construct (Nunnally & Berstein, 1994). Another possibility is that the uncorrelated domains were due to the response bias in this small sample. The next step is to carefully evaluate the dimensions and scale items. This step is critical for determining both observable and latent variables that may not have been thoroughly identified in this analysis. Further, this step is important for confirming the construct of faculty OTE. In sum, the initial evaluation results demonstrated that the FOTES was appropriate and acceptable to measure faculty OTE based upon the panels of experts and the study preliminary sample of the target faculty.

5.1 Lessons learned

The strengths of this educational instrument development are rooted in its conceptual foundation of faculty OTE, good psychometrics, and specific focus on online education. By explicitly measuring TE within online classroom environments, FOTES addresses a significant research gap. Our work has yielded several key insights: FOTES offers a comprehensive evaluation framework that allows for its adaptation to multifaceted contexts, considering teacher actions, personal aspects, work-related factors, and experiential outcomes associated with student learning. This self-evaluation scale assesses teachers' pedagogical expertise in online teaching, encompassing course design, delivery, and the effective utilization of digital technology.

FOTES incorporates the collective perspectives of faculty and students, aiming to improve outcome evaluations, such as virtual student engagement and achievement. This addresses the persistent challenge of ensuring quality in online teaching and learning. Additionally, the instrument emphasizes experiential outcomes, which highlights the need for assessing the impact of teaching on the student learning process, beyond traditional measures such as grades. Our vision is for this instrument to foster a comprehension of teachers' online teaching proficiency across diverse courses and students, determine specific factors influencing outcomes, and provide a base for the ongoing improvement of faculty education and professional growth. The use of this instrument, in turn, will contribute to elevating academic accomplishments and creating richer learning experiences for e-learners worldwide.

5.2 Limitations and future implications

As a result of the funding received for this study, we had an opportunity to pretest the instrument to obtain the preliminary evidence. This pretest, using a small sample, was also an expected study limitation. The scale results in phase III would likely have reflected the overall

performance of the scale items in evaluating FOTE if a larger sample was secured. Because of testing with a small sample size, limited to the public universities of one state in the US, the narrow response ranges indicated less data variance and, thereby, restricted the magnitude and appropriateness of the item properties. Further verification of, and refinement to, the uncorrelated domains and items with the limited single-item reliability is needed. Additionally, the extent to whether the relevant context or content domain is fixed or varying because of other elements (e.g., different degree major, online teaching mode, faculty's digital skills, and technologies used in online teaching, etc.) has not been considered in this preliminary analysis and warrants further exploration.

The study results yielded the explicit and concrete concept of faculty OTE which was important for developing the instrument. Although challenges were encountered during data collection, this tool is scientifically valid and reliable and has a potential to benefit faculty in self-evaluating their teaching effectiveness in online education. Future research with a diverse, large sample is needed to further validate the individual item performance, determine unique item sets within each defined domain considering both analytical and pedagogical approaches, and ultimately enhance the domain competencies to attain the valid measure of the concept with sound psychometric properties.

6. Conclusion

As the global need for online teaching and learning continues to increase, the development of the valid instrument is essential to consistently quantify and tangibly signify faculty OTE as part of quality online education. This is particularly important in light of the current state of limited resources and applications to evaluate OTE for faculty in higher

education. The FOTES, developed in this study, is conceptually based, and has the potential to enhance faculty capability and skill in self-evaluating TE in online education. Although additional psychometric validation is needed, these preliminary results provide empirical evidence to advance future instrument testing. The FOTES instrument is poised to promote outcome evaluation and, in turn, strengthen teaching and learning processes throughout nursing education, research, and practice.

Funding. This study was supported by Sigma Theta Tau International Delta Omega chapter and National League for Nursing/ Sigma Theta Tau International research grants.

Declarations of competing interests. There is no known conflict of interest to be declared by any of the authors.

Acknowledgements. The authors would like to thank all experts who participated in the study panel, completed the reviews, and provided their valuable responses on the newly developed instrument. Our gratitude is given to all the faulty (eligible and ineligible) who expressed their interest in participating in this preliminary evaluation of the study instrument.

Author contributions. YMS and YJC contributed substantially to conception, design, and acquisition of data. All authors were directly involved in data analysis and interpretation. YJC drafted and prepared the manuscript with YMS; all authors have been involved critically revising and editing the paper and approved the final version of the article.

References

Akram, F., Ul Haq, M. A., Malik, H. A., & Mahmood, N. (2021). Effectiveness of online teaching during COVID-19. *International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), 2021,* 568-573. https://doi.org/10.1109/3ICT53449.2021.9582144

Arbaugh, J. B., Cleveland-Innes, M., Diaz, S. R., Garrison, D. R., Ice, P., Richardson, J. C., & Swan, K. P. (2008). Developing a Community of Inquiry instrument: Testing a measure of the Community of Inquiry framework using a multi-institutional sample. *Internet and Higher Education*, *11*(3-4), 133-136. https://doi.org/10.1016/j.iheduc.2008.06.003

Bangert, A. W. (2008). The development and validation of the student evaluation of online teaching effectiveness. *Computers in the Schools, 25*(1), 25-47. https://doi.org/10.1080/07380560802157717

Berk, R. A. (2013a). Top five flashpoints in the assessment of teaching effectiveness. *Medical Teacher, 35* (1), 15–26. https://doi.org/10.3109/0142159X.2012.732247

Berk, R. A. (2013b). Face-to-face versus online course evaluations: A "consumer's guide" to seven strategies. *Journal of Online Learning and Teaching, 9*(1), 140-148.

Chaudhary, A. K., & Israel, G. D. (2014). The Savvy Survey #8: Pilot testing and pretesting questionnaires. *Agricultural Education & Communication*. University of Florida; AEC402.

Chianese, G. (2015). Developing and assessing teaching effectiveness. *Procedia-Social and Behavioral Sciences, 191*, 692-695. https://doi.org/10.1016/j.sbspro.2015.04.218

Chickering, A. W., & Gamson, Z. F. (1987). Seven principles for good practice in undergraduate education. *American Association for Higher Education Bulletin*, *39*(7), 3-7.

Cicchetti, D. V. (1984). On a model for assessing the security of infantile attachment: Issues of observer reliability and validity. *Behavioral and Brain Sciences, 7*, 149-150. https://doi.org/10.1017/S0140525X00026558

Cinches, M. F. C., Russell, R. L. V., Chavez, J. C., & Ortiz, R. O. (2017). Student engagement: Defining teacher effectiveness and teacher engagement. *Journal of Institutional Research South East Asia, 15*(1), 5-19.

Creswell, J. W. (2015). *A concise introduction to mixed methods research*. Sage.

Crews, T. B., Wilkinson, K., & Neill, J. (2015). Principles for good practice in undergraduate education: Effective online course design to assist students' success. *Journal of Online Learning and Teaching, 11*, 87-103.

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research, 5*, 194-197.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Sage.

Doan, S., Schweig, J. D., & Mihaly, K. (2019). The consistency of composite ratings of teacher effectiveness: Evidence from New Mexico. *American Educational Research Journal, 56*(6), 2116-2146. https://doi.org/10.3102/0002831219841369

Elumalai, K.V., Sankar, J. P., Kalaichelvi, R., John, J. A., Menon, N., Alqahtni, M. S. M., & Abumelha, M. A. (2020). Factors affecting the quality of e-learning during the COVID-19 pandemic from the perspective of higher education students. *Journal of Information Technology Education: Research*, *19*, 731-753. https://doi.org/10.28945/4628

Esarey, J., & Valdes, N. (2020). Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment and Evaluation in Higher Education, 45*(8), 1106-1120. https://doi.org/10.1080/02602938.2020.1724875

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). John Wiley & Sons.

Frazer, C., Sullivan, D. H., Weatherspoon, D., & Hussey, L. (2017). Faculty perceptions of online teaching effectiveness and indicators of quality. *Nursing Research and Practice, 2017*, 9374189. https://doi:10.1155/2017/9374189

Grant, J., & Davis, L. L. (1997). Focus on quantitative methods. Selection and use of content experts for instrument development. *Research in Nursing and Health, 20*, 269-274.

Hattie, J. (2015). The applicability of Visible Learning to higher education. *Scholarship of Teaching and Learning in Psychology, 1*(1), 79–91. https://doi.org/10.1037/stl0000021

Islam, N., Beer, M., & Slack, F. (2015). E-learning challenges faced by academics in higher education: A literature review. *Journal of Education and Training Studies, 3*(5), 102-112. https://doi.org/10.11114/jets.v3i5.947

Kang, H. (2012). Training online faculty: A phenomenology study. *International Journal on E-Learning,11*(4), 391-406.

Lauermann, F., & ten Hagen, I. (2021). Do teachers' perceived teaching competence and self-efficacy affect students' academic outcomes? A closer look at student-reported classroom processes and outcomes. *Educational Psychologist, 56*(4), 265-282. https://doi.org/10.1080/00461520.2021.1991355

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563-575.

Layne, L. (2012). Defining effective teaching. *Journal on Excellence in College Teaching, 23*(1), 43-68.

Lederman, D. (2021, September 16). Detailing last fall's online enrollment serge. *Inside Higher Ed.* https://www.insidehighered.com/news/2021/09/16/new-data-offer-sense-how-covid-expanded-online-learning

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382-385. https://doi.org/10.1097/00006199-198611000-00017

Makkonen, R. (2016, September 22). *Using teacher effectiveness data in meaningful ways*. WestEd. https://www.wested.org/rd_alert_online/using-teacher-effectiveness-data-in-meaningful-ways/#

Morgan, J., & Knox, J. E. (1987). Characteristics of 'best' and 'worst' clinical teachers as perceived by university nursing faculty and students. *Journal of Advanced Nursing, 12*, 331-337.

National Center for Education Statistics. (2023). *Undergraduate enrollment.* Condition of Education. U.S. Department of Education, Institute of Education Science (IES). https://nces.ed.gov/programs/coe/indicator/cha

Nunnally, J. C., & Berstein, I. H. (1994). *Psychometric theory*. McGraw-Hill.

Philipsen, B., Tondeur, J., Pynoo, B., Vanslambrouck, S., & Zhu, C. (2019). Examining lived experiences in a professional development program for online teaching: A hermeneutic phenomenological approach. *Australasian Journal of Educational Technology, 35*(5), 46-59. https://doi.org/10.14742/ajet.4469

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Focus on research methods. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing and Health, 30*, 459-467. https://doi.org/10.1002/nur.20199

Price J., Whitlatch, J., Maier, C., Burdi, M., & Peacock, J. (2016). Improving online teaching by using established best classroom teaching practices. *Journal of Continuing Education in Nursing, 47*, 222-227. https://doi.org/10.3928/00220124-20160419-08

Reyes-Fournier, E., Cumella, E. J., March, M., Pederson, J., & Blackman, G. (2020). Development and validation of the online teaching effectiveness scale. *Online Learning, 24*(2), 111-127. https://doi.org/10.24059/olj.v24i2.2071

Ruiz, R., & Sun, J. (2021). Distance education in college: What do we know from IPEDS? *NCES Blog: National Center for Education Statistics.* February 17. https://nces.ed.gov/blogs/nces/post/distance-education-in-college-what-do-we-know-from-ipeds

Singh, V., & Thurman, A. (2019). How many ways can we define online learning? A systematic literature review of definitions of online learning (1988-2018). *American Journal of Distance Education, 33*(4), 289-306. https://doi.org/10.1080/08923647.2019.1663082

Smith, Y. M., Chen, Y. J., & Warner-Stidham, A. (2021). Understanding online teaching effectiveness: Nursing student and faculty perspectives. *Journal of Professional Nursing, 37*(5), 785-794.

Smith, Y. M., & Crowe, A. R. (2017). Nurse educator perceptions of the importance of relationship in online teaching and learning. *Journal of Professional Nursing, 33*(1), 11-19. https://doi.org/10.1016/j.profnurs.2016.06.004

Uttl, B., White, C. A., & Wong Gonzalez, D. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*(2017), 22-42. https://doi.org/10.1016/j.stueduc.2016.08.007

Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2003). Two quantitative approaches for estimating content validity. *Western Journal of Nursing Research, 25*, 508-518. https://doi.org/10.1177/0193945903252998