RESEARCH ARTICLE

Equivalence of entrustable professional activities and context-dependent item sets as summative assessments in undergraduate physical therapy programs

Maria Elizabeth M. Grageda

Author's email address: mmgrageda@up.edu.ph

National Teacher Training Center for the Health Professions, University of the Philippines Manila

ABSTRACT

Background: Summative assessment of student performance should provide information on achievement of program outcomes to support evaluation decisions. Alternative approaches to the traditional assessment systems like the written licensure examinations in Physical Therapy (PT) should be explored to ensure valid measurement of achievement of these terminal outcomes.

Objective: The study aimed at establishing equivalence of two summative assessments new to PT in measuring achievement of the PT outcomes: work-based assessment using Entrustable Professional Activities (EPA) and knowledge-based assessment using Context-Dependent Item Sets (CDIS).

Methodology: Thirty-two newly graduated PT's underwent a one-week EPA assessment and took a 102-item CDIS test (based on 14 clinical vignettes). Qualitative data from blueprint review, group face-to-face interviews with participants and assessors, and field notes from observations, and quantitative data from EPA entrustment decisions and CDIS scores were utilized to ascertain their comparability in terms of Purpose, Administration, Quality and Decisions. This was used to determine the extent of equivalence of the two assessments.

Results: Review of both blueprints show alignment with PT outcomes, with integrative content motivating participants towards professional development. Administration were equally acceptable to users, though EPA had more practice opportunities with a longer assessment time. Entrustment decisions in EPA had a high inter-rater reliability, while CDIS had low reliability, with most items having poor discriminative power. Decisions of "pass" or "fail" had good concordance when high prevalence indices were considered.

Conclusion: There is high extent of equivalence in purpose of EPA and CDIS but are not equivalent in terms of administration. There is moderate equivalence in quality and decisions, with potential for increased concordance if improved quality of CDIS is attained.

Keywords: *Summative assessment, Outcome assessment, Entrustable professional activities (EPA), Context-dependent item sets (CDIS), Comparability, Equivalence*

Introduction

Following outcome-based education (OBE) principles, the Philippine Commission on Higher Education (CHED) issued CMO 55 in 2017: Guidelines for Physical Therapy (PT) Education, outlining 12 outcomes expected of PT graduates. To ensure valid measurement of these outcomes, existing summative assessment systems like traditional written licensure tests are under review, and other assessment formats are being explored. Two summative assessments new to PT were studied to establish equivalence in measuring achievement of terminal outcomes: work-based assessment using Entrustable Professional Activities (EPA) and knowledge-based assessment using Context-Dependent Item Sets (CDIS) in multiple choice question (MCQ) format. Their comparability in terms of Purpose, Quality, Administration, and Decisions were ascertained to determine their extent of equivalence as summative assessments. Results of this study will benefit higher education and regulatory institutions in considering alternative summative assessments like EPA and CDIS in measuring terminal outcomes, based on its relevance and utility within various contexts.

In education, assessment is the process of collection, systematization, and analysis of information about learner

performance useful for evaluation and decision-making [1]. Good quality assessments are characterized by its utility as agents of educational improvement; explicit in values, aims, and standards; relevant, authentic and holistic; reliable and valid; and flexible to learning needs [2,3]. Though formative assessment is emphasized in OBE, summative assessment is essential, considered high stakes for students and highly challenging for educators, involving serious consequences for program improvement, assessment of teaching effectiveness, and program accreditation [3]. Summative assessments measuring achievement of outcomes as a product of learning are outlined in Table 1 based on Miller's pyramid [3].

Table 1. *Common assessment tools grouped by Miller's levels of competence [3]*

Assessment Categories	Description of Assessment	Advantages	Disadvantages
Assessment of "knows" and "knows how"			
MCQs	Selected response with stem, lead-in question, & options. Most common: single-best answer.	Tests broad knowledge & application; easiest to produce statistically reliable results; automated marking	Limited cognitive level tested; not used to extrapolate what students can do; hard to avoid technical flaws; training needed
Short answer	Structured questions, open-ended response. Predetermined model answers.	Easy to create; reasonable content coverage; easier to grade than essays	Require 30-40 questions to match MCQs for reliability; tests same cognitive levels as MCQs but less efficient
Essay/report	Prose in response to stimulus. Points system against a rubric or global rating.	Easy to create; assess written communication skills, complex topics & making coherent arguments	Limited content; modest reliability & interrater agreement; time consuming to grade
Oral exam (viva voce)	One or more examiners ask questions face to face. Blueprinted questions; predetermined rubric.	Traditional; test synthesis under pressure; better used in formative situations	Low reliability; high interrater variation; knowledge level; prone to unconscious biases
Assessment of "shows" (demonstrations of performance in simulated setting)			
Direct observation	Observe in practice: laboratory or clinic; use rating scales	Assess what learners do in real situations; easy to administer; global rating	Requires training; multiple encounters for reliable data
Portfolio	Collection of work samples over time. Goal setting & frequent feedback.	Represent actual performance over time; feedback progress monitoring device	Time consuming; low student acceptance; hard to grade reliably & set standards for high-stakes conditions
Peer assessment	Students assess each other's work using rubric	Encourages student responsibility & ownership; develops judgment; alternative feedback for teamwork & behavior	Grade inflation with less reliability; formative assessment; reluctance to give negative feedback if not anonymous; briefing on assessing giving feedback; should be supervised
Self-assessment	Judgment on own learning, based on established criteria.	Encourages goal setting, responsibility & reflective practice	Grade inflation with less reliability; guided practice to develop self-monitoring skills
360-Degree (multisource feedback)	Survey of individuals within domain of competency, observable behavior & interpersonal skills.	Authentic assessment in real-world setting; multiple perspectives; powerful feedback tool	Reluctance to provide negative feedback; 10 raters for reliable data; difficult to deploy & collect data
Assessment of "is" (consistent demonstration of expected values, attitudes, and behaviors)			
Interviews	Subject-object one-on-one interview; explore professional identity.	In-depth personal exploration	Highly skilled examiner; prerequisite data from "does" level
Standardized survey inventories	New area with limited tools available.	Easy to deploy; theoretically grounded	Self-report; not well validated at this time

CDIS is a variation of the MCQ format, assessing higher cognitive levels by providing real-life context through well-constructed problem-solving stimuli such as: clinical cases or vignettes; diagrams; graphs or tables; patient charts or research reports, followed by several independent MCQs related to the stimuli [4]. This form has established a significant track record in professional licensing and certification examinations in medicine, nursing, and other healthcare occupations despite a sparsity of research on its contribution as an item format [5]. More varied stimulus materials still need to be explored for teaching, administrative, interpersonal, and communication skills [6]. Well-developed CDIS can simulate decision-making skills that could be "the next best thing to being there."

EPA is an approach to assessment operationalizing abstract competencies integrated into relevant and recognizable contexts within the clinical workplace. Assessment is embedded in patient care, integrating assessment data from multiple sources during clinical rotations as trainees perform the various activities. An EPA is a unit of professional practice comprised of discrete tasks fully entrustable to a learner or professional when necessary competencies to execute it unsupervised is demonstrated [7]. Trustworthy performance entails a combination of competencies in a holistic assessment, resulting in summative entrustment decisions to act under a specified level of supervision: 1-No permission to act; 2- Permission to act with direct, proactive supervision in the room; 3-With indirect supervision, not present but is quickly available; 4-Distant supervision not directly available; and 5-Permission to provide supervision to junior trainees. EPAs have been used for both formative and summative assessments in health disciplines like Medicine, Pharmacy, Nursing, and Dentistry and adopted by healthcare education organizations in United States, Netherlands, Canada, Australia, and New Zealand [8,9]. As a relatively new approach to workplace-based assessments, most of the literature on EPAs focus on its conceptual framework and the development process of EPA statements [10,11,12]. This was also reported by O'Dowd *et al*. in their systematic review of 7 years of research on EPA in graduate medical education (2011–2018) after reviewing 49 studies. Implementation and assessment of EPAs, including validation of supervision scales and assessment tools were reported infrequently [13,14,15]. Experience with use of EPAs in post-graduate medical education reveal the need for an adaptive workplace and highly trained faculty as foundations for its success [16]. A study on the application of EPA in a family medicine residency program in Canada report resulting validity (integration of knowledge, skills, attitudes, and values), reliability (multiple assessors evaluating residents over time in different contexts), educational value ("competency scripts" as tools for learning), acceptability to residents and assessors, and cost-effectiveness (assessment as part of existing workplace environment) [17].

Equivalent assessments have equal value or worth to assessors when deciding about learner performance, not necessarily meaning assessments are the same, but equivalent information offered are useful in arriving at judgments [2]. Describing equivalence is achieved by investigating comparability of generic characteristics of assessment: Purpose (For whom? For what?) - the purpose of assessment should be a balance between being a vehicle for learning and measurement and accreditation of learning. Assessments should impact on the level of learning, focusing on high-level and complex thinking [18,19].; Quality (validity, reliability, and usefulness); and how and by whom Administration is conducted and Decisions are made. Decision to pass or fail students are made on the basis of assessment results and is made by an individual or a panel of assessors following a set standard.

As comparisons are diverse and apply to various aspects in assessment such as demands; content of curriculum; student performance; and predictive ability of educational outcomes, identifying a suitable definition of comparability and standards is vital [20]. The purpose or context of comparability should be identified. The generic characteristics of an assessment can be used to define the attributes which serve as grounds for the comparison being made. As comparability is part of validity, these contexts or attributes can be aligned with the five general lines of validity evidence based on the modern idea of validity where evidences should be established that good decisions can be made through meaningful interpretation of scores (Messick's Validity Framework) [19,20]. The context of purpose relates to validity evidence based on content which looks for evidence of a test blueprint and audit that items were prepared according to blueprint. It can also include validity evidence based on consequences of testing which looks at the beneficial or harmful, intended, or unintended impact of the assessment. This can include evidence of improved preparation due to assessment and effects on student motivation and impact to assessors. The context of quality pertains to validity evidence based on internal structure which looks at the reproducibility or reliability of scores. The context of administration relates to validity evidence based on response process which looks at the integrity of the data gathered. Evidence of clear instructions, provision of adequate practice opportunities, and accurate scoring and repository of data contribute to this. Finally, the context of decision relates to validity evidence

based on relationships to other variables which pertain to predictive and concurrent validity [3,21].

Methods to determine comparability include: statistical methods where the 'standard' can be detected and compared through data emerging from assessments; judgmental methods that rely upon human judgement to detect and compare the 'standard' by asking practicing assessors; and survey-observational-anecdotal methods that use information from users through surveys and face-to-face interviews. It is not necessary to show that quality attributes of assessments are identical, but that they are highly similar that any differences in quality attributes have no influence on results.

Methodology

After acquiring exemption from ethical review from the University of the Philippines Manila Research Ethics Board BS PT graduates (2017-2019) from top 10 schools (licensure examination results 2014-2018) located in the National Capital Region were recruited through a faculty member from each institution. Since summative assessment of terminal outcomes is being studied, new graduates with entry-level competencies and minimal experience were sampled. Although the ideal calculated sample size at 95% confidence level and 5% error from approximately 350 graduates is 184, this was not realistic given the design, length of time for EPA assessment, and underlying cost. Recruitment of participants was challenging because graduates were preparing for licensure examination, raising difficulties with availability. A study primer was sent to potential participants, and communication through email, Messenger, or SMS were sent to 60 interested graduates, but only 33 gave their signed written consent. One participant withdrew mid-week, unable to complete assessment. The remaining 32 participants were graduates of seven schools. 72.00% (23) graduated in 2019, 25.00% (8) in 2018 and 3.00% (1) in 2017. 69.00% (22) were licensed, 66.00% (21) had no previous experience in the study site, and 94.00% (30) had no work experience.

Participants were enrolled in the one-week observership program of the PT Section, Department of Rehabilitation Medicine, Philippine General Hospital (PGH), in 7 batches within a span of eight weeks. The program is stable and has been implemented for over ten years. It is designed for PT graduates desiring additional exposure in the hospital setting, with adequate number and variety of patients and experienced supervisors to accommodate 10 enrollees weekly. Participants were assigned to the charity in-patient clinic, handling cases in the wards and specialized care units of the hospital, for bedside and ambulatory care.

The PT Summative EPA Assessment Tool recorded entrustment decisions given by assessors based on level of supervision (1-5) needed by participants to perform the 8 EPAs. Six internal assessors ($\bar{x}$=2.83 years in PGH; $\bar{x}$=2.67 years clinical training) provided ad-hoc entrustment decisions for the 7 batches of participants. They were selected with help from the Chief PT of the research site. Two of the 4 external assessors ($\bar{x}$=13.25 years as PT) conducted assessments on the last day of each batch. They were selected with help from the professional organization. They were given a handbook and a two-hour EPA orientation for levelling off prior to assessments and on the last day of every batch, all internal and external assessors met to deliberate on final entrustment decisions for each participant for each EPA. Concurrent with deliberations, participants took the CDIS, a 2-hour written test with 13 vignettes and 102 MCQs, administered by the research assistant, followed by a semi-structured face-to-face group interview, similar to that conducted with all EPA assessors at the end of 8 weeks (schedule details in Table 2). These interviews were documented through written notes and audio recordings, transcribed verbatim, synthesized, and sent back to the interviewees for member-checking. The researcher observed the assessment processes, including deliberations, documented through written field notes.

Quantitative data (EPA entrustment decisions and CDIS scores) were encoded in MS Excel and SPSS25 for analysis, while qualitative data from interviews and field notes were encoded in MS Word and NVivo10 for thematic analysis. To determine comparability of Purpose, the list of EPAs and the CDIS blueprint were matched with the outcomes to validate content. Group interviews on influence on preparation and motivation were analyzed thematically to determine validity based on consequences. Comparability of Administration was determined through thematic analysis of interviews and field notes regarding the process of assessment. To determine comparability of Quality, validity evidence based on internal structure was analyzed through reliability testing. Interrater reliability (IRR) of EPAs was computed using a two-way mixed, absolute agreement, average-measures intra-class correlation coefficient (ICC) to show the degree of consistency in assessors' entrustment decisions across participants [22]. Reliability of the CDIS was computed using split-half method where scores on each half were correlated to estimate the reliability and statistical correction of the test using Spearman-Brown coefficient and Cronbach's alpha for internal consistency. Item and options analysis determined the quality

Table 2. *Activities experienced by participants during the one-week clinical placement.*

	Activities	Person Conducting the Activity	Schedule
Assessment Activities for EPA	Orientation to the policies of the section, the CHIC, and the study	Internal assessors and the researcher	morning of Day 1
	Screening, assessment, monitoring and treatment of 3-8 patients per day, with documentation through daily PT notes and written evaluations	Participants with varying levels of supervision from internal assessor Observed by researcher	Day 1-5
	Pre- and post-session discussions, individually or in groups for each patient	Assigned internal assessor and participants Observed by researcher	Day 1-5, before the start of the clinic in the morning and in the afternoon in the PT section or in between patients in the wards' nurses' stations during chart reading or while walking to the wards
	Case conference on a selected patient that they evaluated, done in pairs or triads.	Assigned internal assessor and participants Observed by researcher	Day 3-5
	Evidence-based practice (EBP) discussion and journal appraisal where participants develop a clinical question based on the results of their evaluation of an assigned patient, search and select an article to answer their question, present an appraisal, and decide whether results can be applied to the patient	Assigned internal assessor and participants Observed by researcher	Day 3-5
	Teaching activity for a selected group of professionals, students, clients, or caregivers	Participants Observed by researcher and assessors	Day 3, 4 or 5
	Summative assessment	External assessors	Day 5
	Panel meeting	Internal and external assessors Observed by researcher	Day 5
Assessment Activity for CDIS	CDIS examination	Participants Administered by research assistant	Day 5
Study Activities	Face-to-face group interview with participants	Researcher and research assistant	Day 5
	Face-to-face group interview with assessors	Researcher and research assistant	End of 8 weeks

of the test, items, and options. Concordance or degree of agreement between two assessments is often assessed using Cohen's kappa [23]. This was used to determine comparability of Decisions of "pass" or "fail" in EPA and CDIS, and in each PT outcome. Interpretation of data gathered determined the extent of equivalence of the two assessments.

Results

Comparability of Purpose (Content and Consequence)

The two assessments were developed by the same group of experts who used the PT program outcomes as written in CHED CMO 55 as the basis for decisions about the content of both assessments. These were also reviewed by another set of experts through the Delphi technique. Content of EPA and CDIS covered the 12 outcomes, with significant differences in weights (Table 3). Outcomes with higher weights in EPA were low in CDIS and vice versa. Participants and assessors agree that content of both assessments was holistic and integrated, gauging knowledge consistent with PT roles. EPA assessed knowledge, skills and attitudes, integrated within performance of activities, while CDIS assessed cognitive skills across PT roles, integrated within vignettes and questions. The assessments differed in breadth of content.

Table 3. *Distribution of EPAs and CDIS items per PT program outcome*

Program outcome (PO)	Entrustable Professional Activities (EPA)									Context-Dependent Item Set (CDIS)	
	Weight (%)	EPA 1	EPA 2	EPA 3	EPA 4	EPA 5	EPA 6	EPA 7	EPA 8	Weight (%)	CDIS Item no. (Vignette)
Po1 - Apply knowledge of basic sciences	**5.5**			✓	✓	✓	✓		✓	**6.86**	14 (B); 27(C); 37; 38 & 39 (E); 42 & 43(F); 47; 48; 49 & 50 (G); 79 (L)
Po2 - Conduct examination, evaluation & assessment	**10.3**	✓	✓		✓	✓		✓	✓	**8.82**	1 (A); 15 & 25 (B); 28 (C); 37 (E); 42 & 43 (F); 47 (G); 78 & 85 (L); 87 & 91 (M)
Po3 - Demonstrate treatment planning & implementation	**5.8**			✓		✓	✓	✓	✓	**20.10**	2 (A); 16; 25 & 26 (B); 27; 29 & 31(C); 32 & 33 (D); 38; 39 & 41 (E); 44 & 45 (F); 49 & 50 (G); 80; 81; 82 & 84 (L); 86; 89; 90; & 94 (M)
Po4 - Apply teaching-learning principles	**7.5**			✓	✓	✓	✓	✓	✓	**13.23**	11 (A); 17 (B); 31 (C); 36 (D); 45 (F); 51; 52; 53; 54; & 55 (H); 59; 60; 61; 62; & 63 (I)
Po5 - Practice management & leadership skills	**4.5**					✓	✓	✓	✓	**15.69**	3; 6 & 8 (A); 18 & 21(B); 66; 67; 69; 70; 71; 72 & 73 (J); 74; 75; 76 & 77 (K)
Po6 - Demonstrate research-related skills	**13.5**	✓	✓	✓	✓	✓	✓	✓	✓	**10.29**	4 (A); 19 & 24 (B); 30 (C); 40 (E); 46 (F); 56; 57 & 58 (H); 64 & 65 (I)
Po7 - Promote health & improved quality of life	**10.5**		✓	✓	✓	✓	✓	✓	✓	**16.18**	7; 9; 10; 12 & 13 (A); 20 (B); 36 (D); 46 (F); 68 (J); 83 (L); 93 (M); 95; 96; 97; 98; 100; 101; & 102 (N)
Po8 - Engage in lifelong learning	**4.8**			✓			✓	✓	✓	**0.01**	5 (A)
Po9 - Work in interprofessional setting	**13.5**	✓	✓	✓	✓	✓	✓	✓	✓	**2.45**	22 (B); 48 (G); 99 (N)
Po10 - Demonstrate proficient communication skills	**11.5**	✓	✓	✓	✓	✓	✓		✓	**0.05**	35 (D)
Po11 - Demonstrate professional & ethical behaviors	**4.3**				✓	✓	✓		✓	**3.43**	23 & 34 (B); 35 (D); 88 (M)
Po12 - Maximize innovative technology in PT practice	**8.5**	✓		✓	✓	✓	✓		✓	**1.47**	9 (A); 26 (B); 85 (L)
TOTAL	**100.00**									**100.00**	

Note: EPA 1 = Assessing patients seeking PT services; EPA 2 = Screening of clients for PT needs; EPA 3 = Planning & implementing PT plan of care; EPA 4 = Monitoring of PT client outcomes; EPA 5 = Designing, implementing & evaluating PT programs for health & wellness; EPA 6 = Teaching PT concepts & procedures to patients/ clients & their families, students, peers, other health care providers & the public; EPA 7 = Utilizing best available evidence from research for decision-making in PT practice; EPA 8 = Participating in community development activities. CDIS Vignettes: A = Wellness; B = Geriatrics; C = Orthopedics; D = Sports Rehabilitation; E = Musculoskeletal; F & G = Neurologic; H & I = Teaching-learning; J & K = Management & leadership; L = Cardiac Rehabilitation; M = Pediatrics; N = Community Based Rehabilitation (CBR)

Data to ascertain comparability of purpose in terms of consequence of assessment, were mainly from qualitative data through group interviews with both the participants and assessors. Both assessments were perceived to contribute to increased motivation for learning. Four subthemes were identified: Motivation for **self-improvement**. One participant described EPA as a method that can facilitate personal development, while CDIS can prepare them for future experiences. The second sub-theme was **reflection** on their own values to improve weaknesses and fortify strengths for professional development. One participant said "The CS is assessing you, so you also need to assess yourself. You should self-reflect.", pertaining to EPA. Another said "Revisit. Even my own decisions. I asked myself, What is important to me?", pertaining to the reflection process while taking the CDIS exam. Another subtheme under motivation is the provision of **quality**

service, which seemed to be the source of motivation towards excellence for the participants, both for EPA and for CDIS, and not high scores or favorable entrustment decisions. Lastly, confidence in independent decision-making was increased, coupled with added responsibility and **accountability**. EPA encouraged clinical practice while CDIS roused unexplored areas of practice. Reinforced passion and a sense of fulfillment was sensed more in EPA, immediately seeing the impact of PT on patient's lives (Table 4).

EPA had a positive effect on learning for the assessors as clinical educators. It encouraged trust in participants, facilitating rather than directing them towards higher levels of independence, integral to EPA, and an intended outcome for graduates. It encouraged reflection on their own standards, with resolve to help participants mirror themselves as independent PT practitioners. Clinical educators play an important role in modelling, encouraging and supporting learners' demonstrations of integrity, reliability, and humility in addition to ability [24]. It has validated their advocacies and exposed other pressing issues affecting practice.

Participants were oriented to both assessment process, but information provided was not maximized by the participants in preparing for assessment. Majority of the participants did not prepare because of fatigue, lack of time, and belief that previous clinical experiences coupled with "common sense" would be enough, as both assessments were holistic and global. For EPA, they depended on guidance from assessors and for CDIS, past encounters provided the trigger for decision points.

Table 4. *Comparison of effect of the two assessments on participants' motivation to learn and practice PT.*

Sub-themes	EPA	CDIS
Motivation to Learn		
Self-improvement	Maximize opportunities to learn for improvement and continuous growth to be the best version of themselves for the service of patients and not for the grade	To be prepared to face future experiences very similar to those in the cases, independently
	Redemption and overcoming painful past experiences toward personal development	
	Challenge self and consider higher studies for continued professional development	
	Utilize feedback to improve weaknesses and fortify strengths	
Reflection	Reflect on the kind of PT they want to be	Reflect and revisit their own values as basis for their decisions, and the processes involved at arriving at these decisions
	Validation of what they know. Build up confidence, pride in oneself, and the courage to decide independently	Confidence that they know what to do and they can arrive at decisions
Accountability	Exercise freedom and independence and use it for the good of everyone	Accountability to patient and peers/ colleagues
Provision of quality service	Provide quality PT service needed by patients	Strive to be the best PT they can be for the patient
Motivation to Practice		
Setting preference	Practice in a similar setting (hospital) or manage their own clinic, do research, or be effective clinical educators in the future	Explore and learn about other fields and practice pathways aside from the clinical and entertain possibilities to practice in other settings not previously considered
		Widen perspectives as there are multiple ways to approach a problem
Professional worth	Reinforced passion for the profession and fulfillment as a PT, seeing the impact of their work on patient's lives	
	Contribute to improvement of peers	

Comparability of Administration (Response Process)

Clear instructions through an orientation and handbook for EPA and orientation, blueprint, and verbal instructions also written in the booklets and answer sheets for CDIS were provided. Erasures and writing on the booklet were not permitted, preventing changes and marking of important information in the case that could facilitate information processing.

Adequate practice opportunities were evident in the blueprints that included all PT roles and common conditions. EPA excluded those beyond the coverage of rotation, but included intensive care and burn management (acute to semi-acute), not covered in CDIS. EPA offered more practice opportunities: Patient management, Conferences, Teaching Activity, and Institutional Activities. 5-20-minute pre-conferences were opportunities to assess critical thinking and decision-making, important for entrustment. Post conferences provided immediate feedback stimulating reflection and self-assessment. A pattern of decreasing supervision was observed, with internal assessors nearby only with: 1) increased patient anxiety or unstable, complex conditions; 2) unsafe environments putting patient at risk; 3) complex PT techniques, or progression of treatment; or 4)initial encounters with patients, procedures or section equipment. Independence was not always accorded on purpose but as a consequence of patients located in different wards, internal assessor handling multiple participants or multitasking of administrative duties. External assessor supervision was distant, minimizing anxiety of participants.

Fostering relationships inherent to entrustment processes was unique to EPA. Actual work encouraged assessor-participant partnerships with shared responsibilities. EPA assessment was an opportunity for participants to socialize and build new friendships, while assessors collaborated to maximize participant assessment opportunities while addressing patient needs (Figure 1). Unique to CDIS was the impact of the testing environment, especially when it demanded reflection, critical thinking, and decision-making. Taking the test at the end of a tiring week affected participants' overall readiness to perform.

Evidences gathered for entrustment decisions were qualitative and descriptive, represented by an entrustment decision for a specific level of supervision (1-5). Assessors undergo processes of reflection and collaboration during deliberations where consensus was reached through discussions. Assessors' standards for entrustment included four key attributes: minimum competence and entry-level skills ensuring patient safety; empathy; consistency of performance regardless of setting and complexity; and being a reflection of the independent PT clinical practitioner assessors exemplified. For CDIS, results were represented by an overall percentage score (0-100.00%) and a score per outcome based on the blueprint.

Comparability of Quality (Internal Structure)

IRR of EPA showed excellent ICC, except EPA 1 (good ICC) indicating assessors had high degree of agreement and EPAs were rated similarly across assessors. A minimal measurement

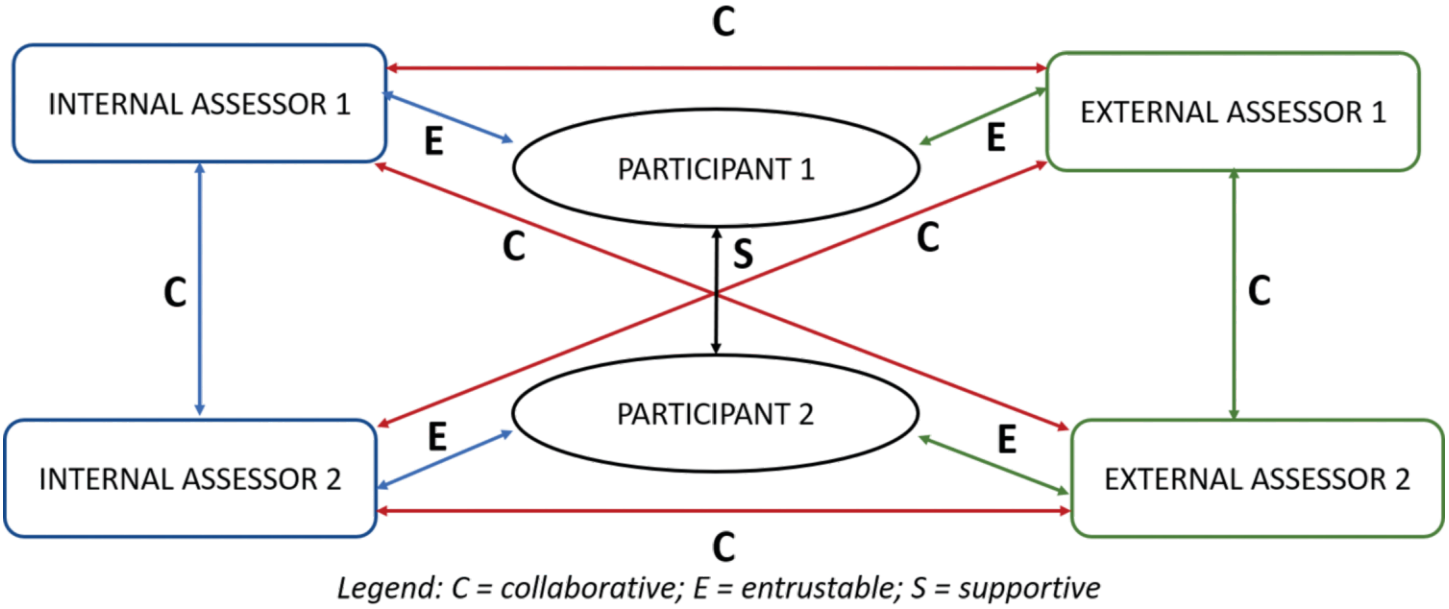

Figure 1. *Relationships developed in EPA assessment*

error was introduced by independent assessors, and statistical power for subsequent analyses is not substantially reduced. Following Classical Test Theory, CDIS showed low reliability (Spearman-Brown coefficient = 00.545) and internal consistency of items (Cronbach's Alpha = 00.331 95% CI [-0.072-0.612]). SEM = 00.04 at sd = 00.05 indicate that almost all score variances is accounted for by error. This was also consistent with results of item-total correlation showing negative values for discrimination at the item level. This may have been brought about by presence of ambiguous or confusing items. Similar responses were expressed by participants during interviews. One participant said, "There are items, maybe, that were like 'Are there no other options?' Like, 'Wait, is the answer really here?' Because it seems the options were limited to one perspective … for me, some choices were not the best answer…".

Item analysis of the 102 MCQs show the lowest p (item difficulty index) value was 0.00 (Items 11, 38, and 43) and the highest was 1.00 (Items 24, 31, and 39). There was good distribution of items with majority having average difficulty, which is desirable when targeting higher order thinking. The lowest D (item discrimination index) was at -0.75 (Item 15), highest was at 0.75 (Items 5 and 49), all with average difficulty. Twenty-three (22.55%) are good items showing average difficulty and satisfactory to high discrimination but had options that were ineffective distracters. There were 3 items (Item 13, 20, and 60) that had good turnout in all options and should be retained. Of the many items with poor discrimination power, those with p=easy had one option rejected by all participants, and those with p=difficult had distracters chosen by more participants over the correct option (Table 5).

Comparability of Decisions (Relation to other Variables)

For EPA, a participant was deemed "pass" if expected entrustment level was reached in ≥50% of the eight EPAs. For CDIS, decisions were based on minimum pass level (MPL) determined by ten experts using the Angoff method. Decisions for each outcome were determined similarly, based on the blueprint.

Table 5. *Distribution of items by difficulty (p) and discrimination (D)*

Discrimination (D)	Difficulty (p)			Total (%)
	Easy p ≥ .76	Average p = 0.25 - 0.75	Difficult p ≤ 0.24	
Poor	9 (8.82)	40 (39.22)	13 (12.75)	62 (60.78)
D ≤ 0.19	*Items 1, 11, 35, 71, 73, 88, 93, 98, 101*	*Items 6, 8, 9, 12, 19, 21, 28, 30, 32, 33, 34, 37, 39, 40, 41, 42, 44, 45, 46, 47, 49, 50, 51, 53, 54, 55, 61, 62, 65, 66, 70. 76, 77, 79, 84, 90, 91, 92, 96, 100*	*Items 15, 16, 22, 23, 24, 26, 31, 38, 57, 68, 72, 89, 102*	
Marginal D= 0.20-0.29	4 (3.92)	9 (8.82)	1 (00.98)	14 (13.73)
	Items 74, 75, 94, 97	*Items 4, 29, 43, 48, 56, 78, 80, 85, 87*	*Item 67*	
Satisfactory D= 0.30 - 0.39	3 (2.94)	9 (8.82)	0	12 (11.76)
	Items 25, 59, 69	*Items 3, 10, 52, 58, 60, 63, 64, 83, 95*		
High D ≥ .40	0	14 (13.73)	0	14 (13.73)
		Items 2, 5, 7, 13, 14, 17, 18, 20, 27, 36, 81, 82, 86, 99		
Total (%)	16 (15.69)	72 (70.59)	14 (13.73)	102 (100.00)

Poor items to discard or revise *Good items to retain*

Overall decisions for EPA and CDIS had no concordance (Cohen's kappa = 0.00), but its interpretation is not straightforward, because of factors that influence its magnitude or interpretation: prevalence, bias, and non-independence of ratings. Prevalence effect, expressed as prevalence index, exists when proportion of agreements on the positive classification (Pass) differs from that of the negative (Fail). Chance agreement is high and kappa is reduced if prevalence of "pass" is either very high or very low (high prevalence index). Bias is the extent to which the decisions disagree on the proportion of positive or negative cases. When a large bias is present, kappa is higher [23]. There was a high prevalence index of 0.81 because of a very low number of "pass" rating in both assessments. Adjusted kappa or Prevalence-adjusted Bias-adjusted kappa (PABAK) resulted in good concordance (PABAK=0.68) showing effects of prevalence and bias, alongside true value of kappa. Only PO12 (Maximizing innovative technology in PT practice) had full concordance while all other outcomes had low concordance. There were high prevalence indices and differences in the maximum attainable kappa (Kmax) and the K value of all outcomes except PO10 (Demonstrate proficient communication skills), showing no concordance, minimal difference in the Kmax and K, and low PABAK. Kmax reflect the extent to which the two assessments' ability to agree on decisions is constrained by pre-existing factors leading to unequal marginal totals such as differences in internal structure and dissimilar sensitivity in measuring specific outcomes.

Discussion

Extent of equivalence of EPA and CDIS as summative assessments

"Equivalence" is used to mean 'a degree of...', or 'extent of...', suggesting that in the real world, equivalence is not absolute as 'equal in value'. Therefore, a subjective measure of what is considered valuable become such only if one attributes a value to it. It is not necessary to show that quality attributes of the two assessments are identical, but that they are highly similar that any differences in quality attributes have no influence on their results. EPA and CDIS are equivalent in purpose, but not in quality, administration, and decisions. In CDIS, low quality of items influenced low performance of participants, but in EPA, performance was still below expected, though showing high reliability. Strengthening their extent of equivalence in terms of decision may be achieved by improving the quality of CDIS. Non-equivalence is considered only in their administration, due to differences in the nature of the assessments.

EPA and CDIS possess characteristics of good assessments, with EPA showing higher utility because of its high reliability. They can be used as summative assessments to measure achievement of outcomes, useful for program improvement, assessment of teaching effectiveness, and accreditation of programs, following the principles of quality assurance and OBE. EPA is a reliable assessment involving multiple assessors and commitment from mature and flexible training institutions. The CDIS, though wanting in quality, can be improved through collaboration with experts in a continuous and iterative process to monitor test and item quality, as its administration still proves to be easier to manage and implement.

Assessment practices in the health professions continuously evolve as the educational milieu continuously reveals new facets to be considered and emphasized. There is no "one best" method. The critic on the quality of written MCQ tests as true reflections of performance has always been an issue in assessment but to date, it remains to be the most widely used assessment format, especially in high stakes, large scale assessments like licensure and certifications [19]. Workplace-based assessment is a unique and indispensable feature of health professions education, yet presents numerous threats to validity and reliability [19]. With the shift to a more outcome-based curricular approach, the CDIS and EPA are presented as two alternative forms or approaches to current traditional written MCQ tests that measure knowledge of disjointed pieces of information and work-based assessments that measure separate specific competencies and skills as listed in checklists. Their extent of equivalence in terms of the 4 attributes can be used as decision points in choosing the more appropriate method within varying contexts (Table 6).

Purpose

A high extent of equivalence of EPA or CDIS as summative assessments render either methods as useful in measuring achievement of PT outcomes. Either of the two assessments can be used based on their alignment with the PT outcomes and their effect on trainee motivation to learn and to practice. Validity evidence based on content is anchored on a test blueprint and audit that it was kept. Blueprints were aligned with PT roles and outcomes, reflected in assessment content described by assessors and participants as holistic. Validity evidence based on consequence looks at effect on motivation to learn and improvements in preparation for assessment. Participants consider being assessed using EPA as an opportunity for them to improve their knowledge, skills, and attitudes with the multiple opportunities to learn. EPA and CDIS motivated participants towards self-improvement and continuous reflection, increasing confidence in providing quality service to patients. According to Reinhard Pekrun's

Table 6. *Decision points on use of EPA or CDIS as summative assessments based on extent of equivalence*

Attribute	Extent of Equivalence	Decision Points	EPA	CDIS
Purpose (as summative assessments)	**High** *Either assessment may be used*	Alignment with PT outcomes	✓ - 8 EPAs and its elaborations can be used	✓- 14 vignettes & high-quality items can be used
		Effect on trainee motivation to learn	✓	✓
		Effect on trainee motivation to practice	✓ - Similar to site	✓ - Broad setting
Decisions	**Moderate** *Choice between EPA or CDIS depends on use & consequence of decision*	Pass/ Fail standards	Pre-determined levels; Useful for overall results, not individual outcomes.	Pre-determined MPL; Useful for overall results, not individual outcomes.
		Decision-making process	Based on panel decision	Based on score
Quality	**Low** *Choice between EPA or CDIS depends on capacity for quality administration*	Reliability	↑ - Ensure reliability through: a. training of assessors; b. multiple assessors; c. panel decision on final supervision level	↓ - Ensure reliability through: a. training item development team; b. adequate test length; c. keeping to blueprint; d. testing environment conducive to thinking
Administration	**None** *Choice between EPA or CDIS depends on assessment context*	Preparation	↓	↑↑
		a. Training	✓ -Training of assessors on EPA; common assessment framework for reliability.	✓ - Training team of item developers on CDIS; item quality for reliability.
		b. Coordination	✓ - With existing stable programs; adequate resources & opportunities to embed EPAs.	✓ - With team of content experts willing to collaborate; suitable testing site.
		Implementation	↑↑	↓
		a. Assessment instructions	✓ - Orientation of trainees to EPA assessment using available media	✓ - Orientation of trainees to CDIS format & blueprint; allow erasures & writing on test booklet
		b. No. of assessors	↑ - Multiple assessors. External ones optional	↓ - At least one proctor during examination
		c. Duration of assessment	↑↑ - Minimum of one week; Dependent on purpose of assessment.	↓ - Minimum of 3 hours for reading of vignettes & analysis
		d. Feedback	↑↑ - Qualitative	↓ - Quantitative
		e. Immediacy of scoring	↓ - Panel decision for summative assessments	↑ - Automated for accuracy.

control–value theory, learners' cognitive appraisal related to perceived control and value of educational activities and outcomes elicit different emotions in learners which influence performance and task outcomes that creates an impact on learners' motivation, learning strategies, cognitive resources, self-regulation and academic achievement [25]. Being given feedback during the process of EPA assessment and on the entrustment decisions and results of the CDIS served as useful information that motivated them to work more to improve identified areas of weakness and fortify their strengths and to continuously challenge themselves. A study by Pitt and Norton identified motivation as one of the 9

dimensions related to how students perceive and respond to feedback[26]. Both positive and negative feedback had a positive motivational effect on students [26] .

Both assessments impressed on accountability of decisions and actions to patients and peers. Entrustment decisions are significant moments of increasing trust and responsibility in trainees aligned with a need for progressive independence or autonomy [7]. EPA and the process of entrustment steered assessors to facilitate independence in participants by providing opportunities for decision-making. Motivating participants and assessors towards self-enrichment shows the educational impact of both assessments on training of future PTs.

The authenticity of both assessments offered by opportunities to work in EPA and stimulation of higher order thinking through real-life scenarios in CDIS, intensified passion to practice as PTs. With EPA, role-modelling by assessors inspired future practice in hospital settings as independent PT clinicians and clinical educators, while CDIS which presented scenarios not limited to clinical settings, inspired practice in health promotion and wellness, and research. Good assessments should influence preparation. Both assessments prompted reflection on past experiences and establishing emotional set, but resulting low performance indicate insufficient preparation.

Administration

Non-equivalent administration stems from different assessment experiences with different requirements for implementation. The choice between EPA or CDIS depends on the assessment context including both time and effort for preparation and implementation of assessment. Validity of evidence based on response process ensure integrity of data throughout the assessment process. Clear instructions were given for both assessments, maximizing different media. Though similar conditions were covered by the assessments there were more opportunities in EPA due to its longer duration with an observed pattern of decreasing supervision from day 1 to day 5. For CDIS, the test was only for 2 hours with participants thinking independently about the vignettes and questions presented. Adequate practice opportunities in CDIS were provided through analyzing cases.

In EPA, actual work gave opportunities to: 1. perform assessment with real patients; 2. be observed during patient interaction, showing professional behaviors aside from technical skills; 3. practice simple to complex skills; 4. receive feedback; and 5. assess skill retention within PT practice context. Case discussions were used to explore professional judgment, assessing higher order thinking; discuss ethical and legal framework of practice; and assess quality of charting and case presentation. Multisource feedback from qualitative observations of assessors, patient satisfaction, audience satisfaction of teaching, and multiple ad-hoc decisions served as basis for entrustment decisions which combine traditional assessment of ability with the right to execute an EPA without supervision, reflecting stepwise acceptance of a trainee as part of the practice community [27]. Entrustment decisions reflect achievement of personal and professional standards, judged by assessors invested in ensuring participants reach high levels of independence and mirror their own trustworthiness as a professional. The entrustable relationship developed between assessor and participant were deemed fundamental to the teaching of core competencies and improving self-confidence [28]. A collaborative relationship was established among assessors, ensuring useful, valid, and reliable data were gathered on participants' performance. A social and supportive relationship developed among participants working together to carry out activities.

Administration includes documentation through a system of checking, scoring, and safekeeping the data. In CDIS, checking was done through Zipgrade application for accurate and immediate results. For EPA, assessors used multisource data during deliberation where individual decisions were discussed before arriving at a consensus. Scores and final decisions were recorded, shared with the participant, kept by the researcher.

Quality

Validity of evidence based on internal structure looks at reliability of scores and reproducibility of assessments. EPA showed high interrater reliability despite challenges in traditional reliability requirements of work-based assessments, because tasks and contexts were varied, including assessors' level and area of expertise and experience, from which judgments were based. “Competence” includes facets not visible in single observations, requiring multiple encounters and raters, demanding longer assessment periods [29]. “What must trainees demonstrate before we can trust them to do the work?” Features of entrustment assessors considered were the same despite variations in experience and expertise resulting in high inter-rater reliability. This framework for assessment followed by EPA assessors addressed threats to validity and reliability and have features similar to those identified by Kennedy *et al.* in [24]. Participants' ability to ensure patient

safety was priority, while competence and positive affective behaviors toward patients for quality PT service was key to full entrustment.

CDIS showed low reliability and internal consistency of items, lacking two major factors that contribute to maximizing reliability of tests: 1) right number and 2) high-quality discriminating items. Intrinsic factors (within test) affecting low reliability include: 1) *Test Length*. The more items, the greater its reliability and vice-versa, but risk of examinee fatigue increases. Reliability testing suggest that the 102 items in the CDIS is insufficient given the breadth of content. Using the Spearman-Brown prophecy formula (Figure 2), lengthening it 2.33 times (adding 136 items) will achieve 0.70 reliability [19,30]; 2) *Discriminative Value (D)*. When items discriminate well between high and low performing groups, item total-correlation and reliability is high and vice-versa. Majority of items in CDIS had poor discrimination, with correct options discarded; and 3) *Instructions*. Complicated and ambiguous directions make understanding questions difficult leading to low reliability. Though CDIS instructions were clear, disallowing erasures decreased participants' chance to change answers and inability to underline important information in the case did not facilitate organization of thought while analyzing problems. Other intrinsic factors: Difficulty Value (p), Item Selection, and Scorer Reliability did not contribute to low reliability of CDIS. Item analysis showed good spread of average difficulty items intended for higher order thinking and though dependent on the cases, items were independent. Scanning of answer sheets gave accurate scores.

Extrinsic factors (outside the test) contributing to low reliability of CDIS include: 1) Group variability. Homogenous ability lowers reliability and vice-versa. Though participants graduated from different educational institutions providing varied educational and clinical experiences, all underwent a curriculum not fully aligned with the PT program outcomes used in the blueprint; 2) Guessing and chance errors. Since items are MCQs with 4 options, there is a 25% chance of answering the items correctly by guessing, which increases error variance and reduces reliability; 3) Environmental conditions. Non-uniform and unfavorable testing environment affected reliability of scores; and 4) Momentary fluctuations. Momentary distractions from the environment, anxiety from pending work, and knowing mistakes cannot be corrected, affected the reliability of CDIS scores.

$$n = \frac{r_{xx}(1 - r_1)}{r_1(1 - r_{xx})}$$

Where r_{xx} = desired reliability
r_1 = obtained reliability and
n = number of times a test is to be lengthened

Figure 2. *Spearman-Brown Prophecy Formula*

There is a lack of consistency in the literature about the effect of sample size on alpha, though there are studies that prove that a small sample size can result to low reliability [31]. This is a consideration since the ideal sample size was not reached. However, sample size estimation for studies that involve Cronbach's alpha test using the formula by Bonett consider three things: the number of items or raters (k), the value of Cronbach's alpha at null hypothesis (CA0) and the expected value of Cronbach's alpha (CA1) [32]. Based on pre-specified alpha at 0.08 and 90.0% power and effect size, a sample size of 11 is needed to obtain the desired value of Cronbach's alpha [32].

Decisions

Low concordance in decisions were due to minimal "pass" ratings in each outcome in both assessments. Differences in internal structure, with high inter-rater reliability of EPA and low reliability, poor discrimination of items and weak performance of options in CDIS served as a pre-existing factor, producing unequal marginal totals. In EPA, high prevalence of "fail" was attributed to limited time to maximize practice opportunities. PABAK and Kmax of CDIS indicate good potential for improvement of internal structure by addressing pre-existing factors related to reliability and quality of items. There was very low concordance in the outcomes PO1 (Application of knowledge of basic sciences), PO4 (Apply teaching-learning principles), PO5 (Practice management and leadership skills), PO7 (Promote health and improved quality of life), PO8 (Engage in lifelong learning), PO10 (Demonstrate proficient communication skills), and PO11 (Demonstrate professional and ethical behaviors) as these are areas not emphasized in the old curriculum that PT educators are not comfortable teaching and assessing.

Conclusion

EPA and CDIS are comparable as summative assessments in terms of purpose, measuring the PT program outcomes and fostering motivation to learn and practice the profession. They are not comparable in terms of quality, with EPA being more reliable than CDIS, and in terms of administration, having very different practice opportunities and scoring processes, but assessments are both acceptable to users. EPA and CDIS are not comparable in terms of

decisions, as each assessment result in different judgments on performance of entry-level professionals.

There is a high extent of equivalence in purpose of EPA and CDIS but are not equivalent in terms of administration. There is moderate equivalence in quality and decisions, with potential for increased concordance between decisions and greater extent of equivalence in these areas if improved quality of CDIS is attained.

Both EPA and DIS are good summative assessments. Either of the two can be used to measure achievement of PT outcomes. However, considering low to moderate extent of equivalence in quality and decisions, the choice between using EPA or CDIS for summative assessment relies on its appropriateness to the assessment context. EPA is more reliable but requires multiple assessors, suitable assessment sites and longer assessment time. CDIS has low reliability, but high-quality items can be obtained through continuous collaborative development and peer review, if a short-duration assessment with uncomplicated administration is preferred.

Acknowledgments

This research was partially supported by a grant from the Philippine Commission on Higher Education (under CMO 04s.2003) and some financial support from the Philippine Physical Therapy Association. Special thanks to the staff and administration of the Physical Therapy Section, Department of Rehabilitation Medicine, Philippine General Hospital for their assistance during implementation of the study.

References

1. Fokienė A, Sajienė L. (2009) Portfolio method in assessment of non-formal and informal learning achievements. The Quality of Higher Education 6:141-158.
2. Murdoch University, Cummings R. (2003) Equivalent assessment: Achievable reality or pipedream?
3. Kibble JD. (2017) Best practices in summative assessment. Advances in Physiology Education 110-119.
4. Tarrant M, Ware J. (2012) A framework for improving the quality of multiple-choice assessments. E-journal of Nurse Educator 37(3):98-104.
5. Oermann MH, Gaberson KB. (2017). Evaluation and Testing in Nursing Education, 5th Ed. New York, NY: Springer Publishing Company, p. 132.
6. Berk RA. (1996) A consumer's guide to multiple-choice item formats that measure complex cognitive outcomes. Pearson Assessment and Information.
7. ten Cate O, Chen HC, Hoff RG, Peters HHB, van der Schaaf M. (2015) Curriculum development for the workplace using entrustable professional activities (EPAs). Medical Teacher 37(11):983-1002.
8. Marotti S, Sim YT, Macolino K, Rowett D. (2018) From assessment of competence to entrustable professional activities. South Australia.
9. Pittenger AL, Chapman SA, Frail CK, Moon JY, Undeberg MR, Orzoff JH. (2016) Entrustable professional activities for pharmacy practice. American Journal of Pharmaceutical Education 80(4):1-4.
10. van Bockel EAP, Walstock PA, van Mook WNKA, *et al.* (2019) Entrustable professional activities (EPAs) for postgraduate competency based intensive care medicine training in the Netherlands: The next step towards excellence in intensive care medicine training. Journal of Critical Care 54:261-267.
11. Chen HC, van den Broek WES, ten Cate O. (2015). The case for use of Entrustable Professional Activities in undergraduate medical education. Academic Medicine 90:431–436.
12. Byrne D, Lydon S, Madden C, O'Dowd E, Boland J, O'Connor P. (2018) The development of entrustable professional activities for the Irish intern year. NUI Galway.
13. Mink RB, Schwartz A, the Steering Committee of the Subspecialty Pediatrics Investigator Network (SPIN), *et al.* (2018). Validity of level of supervision scales for assessing pediatric fellows on the common pediatric subspecialty Entrustable Professional Activities. Academic Medicine. 93:283–291.
14. Schmelter V, März E, Adolf C, *et al.* (2018) Ward rounds in internal medicine: Validation of an Entrustable Professional Activity (EPA) observation checklist. GMS Journal for Medical Education 35(2):Doc17.
15. O'Dowd E, Lydon S, O'Connor P, Madden C, Byrne D. (2019) A systematic review of 7 years of research on entrustable professional activities in graduate medical education, 2011–2018. Medical Education 53:234–249.
16. Van Loon KE, Driessen EW, Teunissen PW, Scheele F. (2014) Experiences with EPAs, potential benefits and pitfalls. Medical Teacher 36:698-702.
17. Schultz K, Griffiths J, Lacasse M. (2015) The application of Entrustable Professional Activities to inform competency decisions in a family medicine residency program. Academic Medicine 90:888–897.
18. Archer E. (2017) The assessment purpose triangle:

Balancing the purposes of educational assessment. Frontiers in Education 2(41)(5):1-6.
19. Price M, Carroll J, O'Donovan B, Rust C. (2011) If I was going there I wouldn't start from here: a critical commentary on current assessment practice. Assessment and Evaluation in Higher Education 36(4):479–492.
20. Elliott G. (2013) A guide to comparability terminology and methods. Cambridge Assessment.
21. Yudkowsky R, Park YS, Downing SM. (2020) Assessment in health professions education. NY: Routledge.
22. Mc Graw KO. (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods 1(1):30-46.
23. Kwiecien R, Kopp-Schneider A, Blettner M. (2011) Concordance analysis. Deutsches Ärzteblatt International 108(30):515–521.
24. Chen HC, ten Cate O. (2018) Assessment through entrustable professional activities. In: Delany C, Molloy E (eds.). Learning and Teaching in Clinical Contexts: A Practical Guide, Australia: Elsevier.
25. Gomez-Garibello C, Young M. (2018) Emotions and assessment: considerations for rater-based judgements of entrustment. Medical Education 52(3):254-262.
26. Pitt E, Norton L. (2017) 'Now that's the feedback I want!' Students' reactions to feedback on graded work and what they do with it. Assessment & Evaluation in Higher Education 42(4):499–516
27. ten Cate O. (2016) Entrustment as assessment: Recognizing the ability, the right, and the duty to act. Journal of Graduate Medical Education: 261-262.
28. Jackson D, Davison I, Adams R, Edordu A, Picton A. (2019) A systematic review of supervisory relationships in general practitioner training. Medical Education 53(9):874-885.
29. Pangaro L, ten Cate O. (2013) Frameworks for learner assessment in medicine. Medical Teacher 35(6):e1197-e1210.
30. Brown JD. (2001) Can we use the Spearman-Brown prophecy formula to defend low reliability? Shiken: JALT Testing & Evaluation SIG Newsletter 4(3):7 -11.
31. Abdelmoula M, Chakroun W, Akrout F. (2015) The effect of sample size and the number of items on reliability coefficients: alpha and rhô: A meta-analysis. International Journal of Numerical Methods and Applications 13(1):1-20.
32. Mohamad AB, Evi DO, Nur AB. (2018) A review on sample size determination for Cronbach's alpha test: a simple guide for researchers. Malaysian Journal of Medical Sciences 25(6):85–99.