Discussant: Rachel Yudkowsky, MD, MHPE
Facilitator: To Be Determined

Combining Scores Based on Compensatory and Noncompensatory Scoring Rules to Assess Resident Readiness for Unsupervised Practice: Implications From a National Primary Care Certification Examination in Japan

Hirotaka Onishi, MD, MHPE, PhD, Yoon Soo Park, PhD, Ryo Takayanagi, MD, and Yasuki Fujinuma, MD

Abstract

Purpose

Competence decisions in health professions education require combining scores from multiple sources and identifying pass–fail decisions based on *noncompensatory* (required to pass all subcomponents) and *compensatory* scoring decisions. This study investigates consequences of combining scores, reliability, and implications for validity using a national examination with subcomponent assessments.

Method

National data were used from three years (2015, 2016, and 2017) of the Japan Primary Care Association Board Certification Examination, with four subcomponent assessments: Clinical Skills Assessment–Integrated Clinical Encounter (CSA-ICE), CSA–Communication and Interpersonal Skills (CSA-CIS), Multiple-Choice Questions (MCQ), and Portfolio. Generalizability theory was used to estimate variance components and reliability. Kane's composite reliability and kappa decision consistency were used to examine the impact of using compensatory and noncompensatory scoring.

Results

Mean performance (n = 251) on the CSA-ICE, CSA-CIS, MCQ, and Portfolio subcomponent assessments were, respectively, 61% (SD = 11%), 67% (SD = 13%), 74% (SD = 8%), and 65% (SD = 9%); component-specific Φ-coefficient reliability ranged between, respectively, 0.57 and 0.67; 0.50 and 0.60; 0.65 and 0.76; and 0.87 and 0.89. Using a completely noncompensatory scoring approach on all four subcomponents, decision-consistency reliability was 0.33. Fully compensatory scoring yielded reliability of 0.86.

Conclusions

Assessing a range of abilities in making entrustment decisions requires considering the balance of assessment tools measuring distinct but related competencies. These results indicate that noncompensatory pass–fail decision making, which seems more congruent with competency-based education, may lead to much lower reliability than compensatory decision making when several assessment subcomponents are used.

In health professions education, the competence of junior physicians and the associated entrustment decisions leading to unsupervised practice can be measured by combining information from multiple sources.[1–3] This is often conducted in the context of combining trainee performance from related, but unique assessments.[4,5] For example, performance on multiple-choice tests and standardized patient (SP) encounters can be combined to form composite scores that reflect competencies in clinical knowledge and patient care skills from both assessments.[6,7] Prior studies on composite scores have addressed issues related to identifying weights of each assessment, by balancing their clinical or curricular relevance with their measurement characteristics such as reliability and interassessment correlation. Moreover, studies have also made progress in discussing the reliability and validity of composite scores, identifying factors that contribute to their validity evidence.[1,4] However, an area that still requires much needed attention, particularly in the context of competency-based medical education (CBME), is the discussion between *compensatory* and *noncompensatory* scoring of multicomponent assessments. This issue focuses on whether trainees should pass all subcomponents to pass the entire assessment (i.e., noncompensatory), or whether performance on one subcomponent can compensate for performance on other sections of the assessment (compensatory).

The issue of compensatory and noncompensatory scoring requires discussion of the assessment challenges embedded in making *composite decisions*.[8,9] For example, if board certification examinations consist of multicomponent measures—(1) medical knowledge (MK), (2) patient history taking and physical examination (H&P), and (3) communication and interpersonal skills (CIS)—test developers and professional societies need to decide whether examinees must pass all individual subcomponents separately to pass the entire test (e.g., pass all three subcomponents), or whether parts of the examination can be combined to compensate for performance on other components (e.g., specifying H&P and CIS as compensatory). Engaging in a deeper discussion around compensatory and noncompensatory consequences of assessment scores and examinee pass–fail results addresses a fundamental aspect of consequential validity.[10,11]

Please see the end of this article for information about the authors.

Correspondence should be addressed to Hirotaka Onishi, International Research Center for Medical Education, Graduate School of Medicine, The University of Tokyo, 2F Igakubu-Sogochuokan, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan 113-0033; telephone: (+81) 3-5841-3534; e-mail: onishi-hirotaka@umin.ac.jp.

Acad Med. 2018;93:S45–S51.
doi: 10.1097/ACM.0000000000002380

Composite and noncompensatory subcomponent scores are prevalent in local medical school and residency assessments and also in high-stakes licensing examinations and board certification examinations. For example, the United States Medical Licensing Examination Step 2 Clinical Skills (CS) examination requires examinees to pass the Integrated Clinical Encounter (ICE), Communication and Interpersonal Skills (CIS), and Spoken English Proficiency subcomponents separately, in a noncompensatory manner, to pass the entire Step 2 CS examination.[12] The ICE subcomponent uses a compensatory scoring approach, combining scores from the physical examination skills based on the SP encounter and the diagnostic justification skills based on the examinee's patient note assessment.[4] As such, the Step 2 CS examination includes both compensatory and noncompensatory components in determining the overall examinee performance. Likewise, board certification examinations administered by medical specialty boards in the United States are also based on the principle of noncompensatory scoring. The American Board of Surgery Initial Board Certification examination requires examinees to pass both the MK component (qualifying examination), measured using multiple-choice questions; and an oral examination component (certification examination), in a noncompensatory manner, to achieve board certification.[13]

There are several reasons that support noncompensatory scoring, including considerations for patient safety, proficiency in distinct competencies (e.g., Accreditation Council for Graduate Medical Education or CanMEDS), and curricular relevance, among others.[14–16] However, noncompensatory scoring also has direct implications on composite score reliability and diminishing psychometric properties of the assessment, for every incrementally added noncompensatory assessment measure.[17–20] As such, a comprehensive analysis of these considerations can benefit how test scores are used and interpreted.

In this study, we use national data from three consecutive years to examine internal structure and consequential validity evidence of composite scores and composite decisions related to compensatory and noncompensatory scoring. We use the Board Certification Examination for family medicine specialists in Japan to contribute to this empirical discussion. Implications for weighting of assessment components and pass–fail consequences are discussed.

Method

Study context and participants: Japan

In Japan, each clinical academic society or association has been independently managing board certification for different physician specialties. For primary care specialties, including family medicine, the Japan Primary Care Association (JPCA) manages the certification examination. The JPCA was established in 2010 by the merger of three academic societies: Japanese Academy of Family Medicine, Japanese Society of General Medicine, and Japanese Medical Society of Primary Care. The JPCA is the administering body responsible for developing, scoring, and pass–fail decision making for family medicine specialists in Japan, who must complete both two-year postgraduate clinical training and three-year family medicine training, similar to the system in the United Kingdom.

Data for this study come from three consecutive JPCA examinations, administered nationally to all family medicine specialists in Japan. From 2015 to 2017, 67, 79 and 105 examinees took the examination, respectively. The results of this study are essential to improving the quality of examinations, as Japan is also transforming the board certification system by centralizing them into the Japanese Medical Specialty Board in subsequent years to promote increased public accountability, with 19 specialist program certifications in fundamental areas to be managed through a central board certification system.

Assessment components

The JPCA examination consists of four subcomponents: (1) Clinical Skills Assessment–Integrated Clinical Encounter (CSA-ICE), which comprehensively measures history-taking and physical examination skills, clinical reasoning, interpersonal and communication skills, and professionalism of candidates scored by two independent physician examiners, as candidates rotate through six different SP encounters; (2) Clinical Skills Assessment–Communication and Interpersonal Skills (CSA-CIS), which measures patient-centered interpersonal and communication skills using ratings from SPs; (3) Multiple-Choice Question (MCQ) test measuring clinical knowledge; and (4) Portfolio assessment, which measures integrative clinical understanding of patient care, system-based practice, and practice-based learning and improvement. JPCA discloses the use of noncompensatory pass–fail decision making.

The CSA-ICE and CSA-CIS are measured as parts of a six-station CSA. Portfolio consists of reports for 18 different areas. Its scoring rubric for each area is disclosed prior to test administration to all examination candidates. All candidates are required to select one case with their best performances for each area. Two independent physician raters assess reports in each item. For the MCQ, each test administration year used slightly different test development procedures, and as such, the numbers of MCQ items increased (2015: 41 items; 2016: 94 items; 2017: 103 items) because part of the format was a previously modified essay question. Psychometric analyses of each assessment component are routinely conducted, including item analysis and reliability estimation to evaluate their measurement characteristics.

Analysis

Data were compiled across the three testing years (2015, 2016, and 2017) and analyzed using the following steps. First, we independently examined the descriptive statistics (item and total score distribution) and reliability of the four components: (1) CSA-ICE, (2) CSA-CIS, (3) MCQ, and (4) Portfolio. Subsequently, we applied component-specific weights and reliability using the Kane approach to calculate composite scores and their reliability estimates.[4,10] Traditional approaches for estimating reliability or composite scores may ignore the reliability of each assessment and their corresponding pairwise associations. For example, if an assessment system includes components of high and low reliability that are moderately correlated, simply taking their average may not take into account the nuanced measurement characteristics of each assessment. As such, under the Kane approach, the reliability of composite scores can be

expressed as a combination of reliability coefficients of component and correlation coefficients, thereby generating more psychometrically sound composite scores and reliability estimates. Finally, we used component-specific passing standards specified by the JPCA committee to derive pass–fail rates by treating each component as noncompensatory and also using a compensatory approach. The final step was conducted to simulate consequences of reliability and pass–fail results when varying scoring approaches were used.

Generalizability study (G study) was used to estimate variance components and to estimate reliability for each test component by testing year.[21,22] CSA-ICE used a three-facet design, with person (*p*: examinees) crossed by raters (*r*) and items (*i*) nested in station (*s*), $p \times [(r \times i)\colon s]$. For the CSA-CIS and MCQ, a one-facet G study design was used, $p \times s$ and $p \times i$, respectively. Variance components for Portfolio were estimated using a two-facet G study design, $p \times (r\colon i)$. Variance components were estimated using urGENOVA. Reliability indices for component scores were based on the Φ-coefficient, as all examinations were conducted as criterion-referenced assessments.

Next, pass–fail rates and reliability were calculated by both noncompensatory and compensatory scores for three consecutive years. This calculation was based on the actual results but only a simulation because JPCA does not disclose detailed information regarding its cut scores and standard-setting procedures.

To examine the consequences of decision-consistency reliability based on compensatory and noncompensatory scoring, we derived reliability estimates for three scenarios:

1. Four noncompensatory subcomponents: assuming noncompensatory scoring for all four test subcomponents (CSA-ICE, CSA-CIS, MCQ, and Portfolio);
2. Three noncompensatory subcomponents: assuming noncompensatory scoring for three test subcomponents, combining the CSA-ICE and CSA-CIS subcomponents (CSA, MCQ, and Portfolio); and
3. Fully compensatory: using a fully compensatory scoring approach combining scores from all four subcomponents.

The weights for CSA-ICE, CSA-CIS, MCQ, and Portfolio were initially set as 20%, 15%, 30%, and 35%, respectively, following faculty consensus. Then, the weight for CSA-ICE was altered under the proportional weights of three other assessment subcomponents (CSA-CIS, MCQ, and Portfolio). These scoring decisions were based on discussion with the JPCA committee on possible score combinations. Noncompensatory score reliability was estimated using decision consistency and *k*-way kappa statistics.[23–25] Compensatory composite score reliability was estimated using the Kane method.[10]

Data compilation and analyses were conducted using Stata 14 (Stata Corp, College Station, Texas). This study was approved by the institutional review board at the University of Illinois at Chicago.

Results

Descriptive statistics

A total of 251 examinees took the JPCA, across the four examination components: CSA-ICE, CSA-CIS, MCQ, and Portfolio (2015: n = 67; 2016: n = 79; 2017: n = 105). The number of OCSE stations in the CSA was fixed to 6 stations; the Portfolio was also fixed to 18 items. The number of MCQ items increased between testing years to reflect changes in the examination format. Table 1 shows the numbers of items for each test component, numbers of examinees of each year, and the total test score distribution. Overall, the mean scores for CSA-ICE, CSA-CIS, MCQ, and Portfolio components across years were 61% (SD = 11%), 67% (SD = 13%), 74% (SD = 8%), and 65% (SD = 9%), respectively.

Variance components, interassessment correlation, and reliability

Variance components. Table 2 shows variance components and percent variance component for each assessment facet. For the CSA-ICE, person variance ranged between 5% and 12%, with the largest variance component due to person-station of 15% to 20%, indicating case specificity. For the CSA-CIS, person variance ranged between 13% and 20%. For the MCQ, person variance ranged between 2% and 3%, with majority of variance due to items, 21% to 28%. Finally, for the Portfolio component, the person variance ranged between 17% and 20%; person–item interaction accounted for 20% to 22% of total variance, also indicating item specificity.

Interassessment correlation and reliability. Correlations between the

Table 1
Distribution of Component Scores by Year: Descriptive Statistics

		CSA-ICE[a]		CSA-CIS[b]		MCQ[c]		Portfolio[d]	
Year	No. of examinees	No. of stations	Mean (SD)	No. of stations	Mean (SD)	No. of items	Mean (SD)	No. of items	Mean (SD)
2015	67	6	64.1 (12.6)	6	64.3 (14.9)	41	81.0 (7.2)	18	71.8 (7.5)
2016	79	6	55.9 (11.1)	6	67.6 (12.4)	94	71.9 (7.6)	18	60.4 (10.3)
2017	105	6	62.5 (8.6)	6	69.0 (11.8)	103	67.9 (8.3)	18	63.9 (9.3)
Overall[e]	251	—	60.8 (10.7)	—	67.3 (13.0)	—	73.6 (7.7)	—	65.4 (9.0)

[a]CSA-ICE measures integrative skills of history taking, physical examination, clinical reasoning, interpersonal and communication, and professionalism of candidates as they rotate through different standardized patient encounters. This is measured by two independent physician examiners.
[b]CSA-CIS measures interpersonal and communication skills using ratings from standardized patients.
[c]MCQ is an assessment of clinical knowledge using multiple-choice questions.
[d]Portfolio assessment measures integrative clinical understanding including patient care, system-based practice, and practice-based learning and improvement.
[e]*Overall* takes the descriptive statistics from accumulated data across the years.

Table 2
Generalizability Theory: Variance Components by Assessment

		2015			2016			2017		
Examination	**Effect**	***df***	**VC**	**% VC**	***df***	**VC**	**% VC**	***df***	**VC**	**% VC**
CSA-ICE	person (*p*)	66	1.069	11.9%	78	0.075	8.5%	104	0.039	4.6%
	station (*s*)	5	0.096	1.1%	5	0.043	4.9%	5	0.097	11.3%
	item: station (*i*: *s*)	18	0.000	0.0%	33	0.011	1.2%	26	0.000	0.0%
	rater (*r*)	1	0.201	2.2%	1	0.000	0.0%	1	0.000	0.0%
	p × *s*	330	1.597	17.8%	390	0.175	19.9%	520	0.125	14.6%
	p × *i*: *s*	1,188	1.881	21.0%	2,574	0.059	6.7%	2,704	0.000	0.0%
	p × *r*	66	0.138	1.5%	78	0.000	0.0%	104	0.001	0.1%
	s × *r*	5	0.027	0.3%	5	0.000	0.0%	5	0.000	0.0%
	i × *r*: *s*	18	0.283	3.2%	33	0.092	10.5%	26	0.099	11.6%
	p × s × *r*	330	0.382	4.3%	390	0.000	0.0%	520	0.000	0.0%
	p × *i* × *r*: *s*, error	1,188	3.302	36.8%	2,574	0.425	48.3%	2,704	0.497	57.9%
	Φ-coefficient reliability		0.68			0.67			0.57	
CSA-CIS	person (*p*)	66	1.098	13.7%	78	0.884	12.7%	104	0.135	19.9%
	station (*s*)	5	0.185	2.3%	5	0.938	13.5%	5	0.015	2.2%
	p × *s*, error	330	6.728	84.0%	390	5.154	73.9%	520	0.530	78.0%
	Φ-coefficient reliability		0.50			0.51			0.60	
MCQ	person (*p*)	66	0.003	3.3%	78	0.003	1.5%	104	0.005	2.4%
	item (*i*)	40	0.024	22.9%	93	0.051	28.0%	102	0.046	21.0%
	p × *i*, error	2,640	0.077	73.8%	7,254	0.129	70.5%	10,608	0.167	76.6%
	Φ-coefficient reliability	0.65	0.65			0.67			0.76	
Portfolio	person (*p*)	66	48.486	17.5%	78	3.038	19.9%	104	2.442	16.9%
	item (*i*)	17	24.794	9.0%	17	0.950	6.2%	17	0.483	3.3%
	rater: item (*r*: *i*)	18	6.735	2.4%	18	0.966	6.3%	18	0.989	6.8%
	p × *i*	1,122	60.259	21.8%	1,326	3.249	21.2%	1,768	2.884	19.9%
	p × *r*: *i*, error	1,188	136.051	49.2%	1,404	7.095	46.4%	1,872	7.676	53.0%
	Φ-coefficient reliability		0.87			0.89			0.87	

Abbreviations: CSA-CIS indicates Clinical Skills Assessment–Communication and Interpersonal Skills; CSA-ICE, Clinical Skills Assessment–Integrated Clinical Encounter; MCQ, Multiple-Choice Questions; VC, variance component.

CSA-ICE and CSA-CIS were 0.62. Correlations between CSA-ICE and other assessments ranged between 0.23 and 0.39. For CSA-CIS, interassessment correlations with other assessments were lower, at 0.10. Correlations between MCQ and Portfolio were 0.33. The overall Φ-coefficients for CSA-ICE, CSA-CIS, MCQ, and Portfolio ranged between, respectively, 0.57 and 0.68, 0.50 and 0.60, 0.65 and 0.76, and 0.87 and 0.89.

Compensatory versus noncompensatory scoring

Impact of composite scores. Figure 1 shows the plot of composite scores using the Kane method, using a fully compensatory approach combining scores across all four components. Weights used to combine the scores were, respectively, 20% 15%, 30%, and 35% for CSA-ICE, CSA-CIS, MCQ, and Portfolio. These weights were selected by JPCA members to reflect content and clinical relevance based on their testing blueprint. The x-axis shows the weighted mean, simply taking the sum of scores using the weights (= [20% × CSA-ICE] + [15% × CSA-CIS] + [30% × MCQ] + [35% × Portfolio]). The Kane composite scores on the y-axis were based on the Kane method, which uses the weighted mean but also takes into consideration the individual assessment reliability, interassessment correlation, and weights. The scatter plot based on the composite scores is not a linear line, indicating that there are variations in the score range. For example, for a weighted mean score of 60%, there are seven different Kane composite scores (range 23–33). As such, depending on where individual pass–fail scores were assigned using a weighted mean approach, the composite score mean could yield a different score, leading to varying pass–fail decisions.

Figure 2 shows the relationship between varying weights of CSA-ICE and its impact on overall composite score reliability. The figure shows a parabolic relationship, with reliability maximized when CSA-ICE is weighted at 15%.

Using a completely compensatory scoring approach (combining scores across the four subcomponents) yields a composite score reliability of 0.86. This is based on individual reliability estimates of 0.64, 0.54, 0.69, and 0.88 for the CSA-ICE, CSA-CIS, MCQ, and Portfolio, respectively. Table 3 shows the results.

Noncompensatory scores: Assuming four subcomponents. Table 3 shows the decision-consistency kappa reliability for different noncompensatory scoring scenarios. When noncompensatory scoring was used for the four

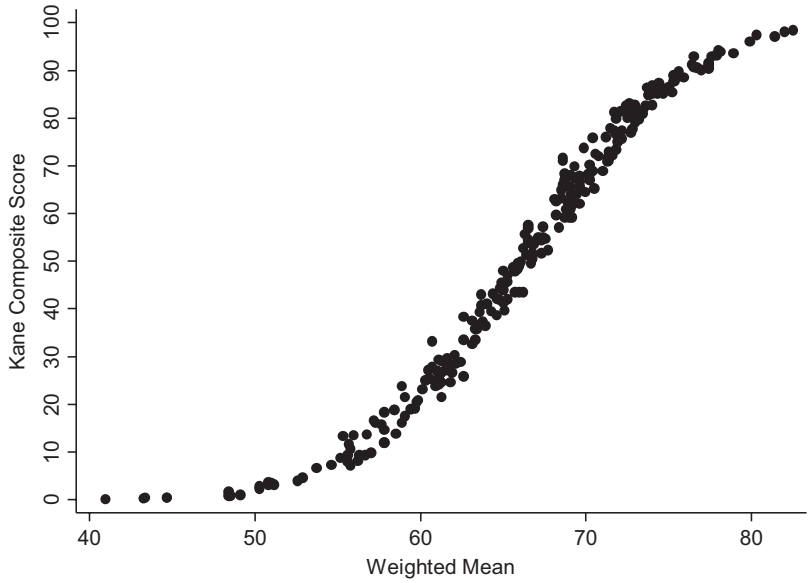

Figure 1 Scatter plot of weighted mean and Kane composite scores (n = 251).

subcomponent assessments (i.e., examinees must pass CSA-ICE, CSA-CIS, MCQ, and Portfolio separately to pass the entire examination), the decision-consistency reliability was 0.33. This is in contrast to a composite score reliability of 0.86 when all subcomponents were assumed to be compensatory.

Noncompensatory scores: Assuming three subcomponents. In our next simulation, we combined the CSA assessments (CIS-ICE and CIS-CIS). Then, using the three subcomponent scores (CSA, MCQ, and Portfolio), we calculated the decision-consistency reliability, which increased from 0.33 (four noncompensatory subcomponents) to 0.50.

When the three subcomponent scores were combined to form a compensatory composite measure, the composite score reliability only had a minor change from 0.86 (compensatory composite score created from four subcomponents) to 0.87 (compensatory composite score created from three subcomponents).

Discussion

This study contributes to the broad discussion on composite scores and consequences of using compensatory and noncompensatory pass–fail decision making. Our study reports that using noncompensatory scoring has direct consequences on composite reliability and decision consistency, which may be mitigated by combining the compensatory approach within the noncompensatory components of the assessment. Prior discussions in the literature have focused on using weights to combine scores, without a deeper discussion on how compensatory or noncompensatory scoring could affect the reliability and overall validity of the assessment.[1,4] In the CBME framework, where learners are required to master and excel in unique but distinct competencies, our findings can form a useful basis for further discussion.[26]

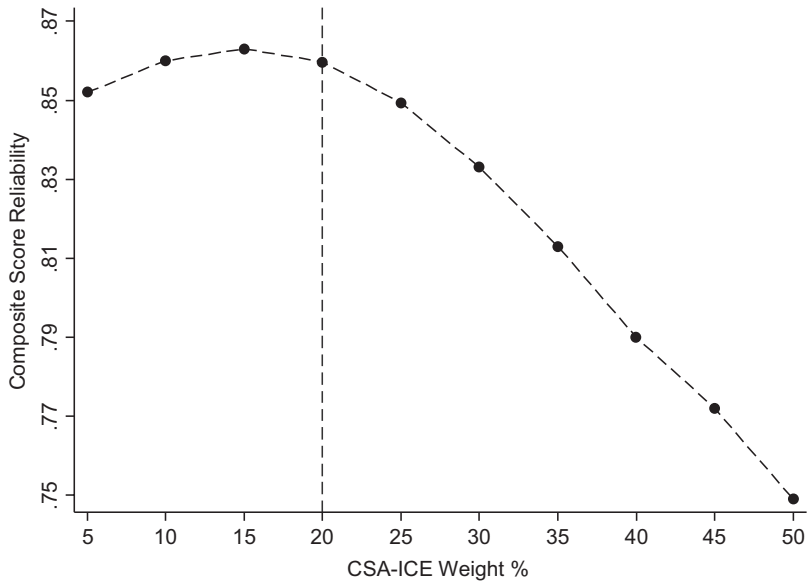

Figure 2 Relationship between CSA-ICE and CSA-CIS weight and composite score reliability. Note: Vertical line inserted at 20%, which indicates the initial weight for CSA-ICE.

Our findings can be largely divided into two parts. First, we provide important implications on compensatory scoring, based on a composite score approach. Competencies and entrustment decisions in medical education require combining information from various sources, including knowledge assessments, workplace-based assessments, and skill assessments such as SP encounters. Findings from this study provide insights on how weights can be specified. For example, our results in Figure 2 show a parabolic relationship between weights and composite score reliability when scores are combined. Such information could be shared with assessment committees or with clinical competence committees (CCCs) in identifying optimal weights that can maximize reliability, validity, and the overall curriculum focus of the educational program. Such information can also be used to identify deficiencies in the program with respect to developing additional assessment tools that can advance decisions for promotion.

Our study also provides some empirical guidance on how noncompensatory scoring can impact the psychometric consequences of the assessment. Prior studies by Hambleton[27] have shown that increasing the number of noncompensatory measures can reduce the overall reliability. A well-known example used in the measurement literature is an assessment with five subcomponent assessments scored in a noncompensatory manner.

Table 3

Pass–Fail Rates and Reliability: Consequences of Noncompensatory and Compensatory Scoring

		Four components[a]		Three components[b]	
Pass–fail type	Assessment	Fail %	Reliability	Fail %	Reliability
Noncompensatory	CSA-ICE	7.57	0.64	9.56	0.75
	CSA-CIS	6.37	0.54		
	MCQ	4.38	0.69	4.38	0.69
	Portfolio	6.77	0.88	6.77	0.88
	Overall (non-compensatory)	18.73[d]	0.33[e]	18.73[d]	0.50[e]
Compensatory[c]	—	4.78[d]	0.86[e]	7.57[d]	0.87[e]

Abbreviations: CSA-CIS indicates Clinical Skills Assessment–Communication and Interpersonal Skills; CSA-ICE, Clinical Skills Assessment–Integrated Clinical Encounter; MCQ, Multiple-Choice Questions.
[a]Four components of the assessment are CSA-ICE, CSA-CIS, MCQ, and Portfolio.
[b]Three components of the assessment are CSA (combining ICE and CIS), MCQ, and Portfolio.
[c]Interassessment correlations range between 0.10 and 0.62. These were used to estimate the Kane composite score reliability.
[d]Overall fail rate based on noncompensatory and compensatory for four- and three-component scoring methods, respectively.
[e]Reliability estimates corresponding to overall fail rates.

Even if each subcomponent has a reliability of 0.90, the overall decision-consistency reliability can drop to 0.59 ($= 0.90 \times 0.90 \times 0.90 \times 0.90 \times 0.90$). This example, of course, does not take into account the interdependence between the assessments (assuming that each assessment is fully independent). However, it conveys the straightforward message that noncompensatory scoring has its benefits, but also its costs on significantly reducing the overall decision-consistency reliability. In other words, even if a well-prepared learner were to take the examination, the overall reliability would be 0.59, adding to uncertainty of whether the learner failed the examination because of his or her ability, or based on the noncompensatory nature of the test.

In our study, we simulated data to show that an assessment with four subcomponent individual reliability estimates ranging between 0.54 and 0.88 would yield a decision-consistency reliability of 0.33—this is much lower than a fully compensatory approach composite score reliability of 0.86. We also show that by simply combining two of the subcomponents, the reliability could increase from 0.33 to 0.50. These empirical results could shed light into how programs and testing organizations may develop test-scoring procedures and identify pass–fail decisions. In this process, we are not explicitly stating that noncompensatory scoring should be avoided. Noncompensatory scoring does have its benefits to ensure that learners are fully competent in specific content areas, which can be important for ensuring patient safety concerns. However, our findings do underscore the need to consider both psychometric consequences, in addition to patient safety factors that can lead to making the compensatory or noncompensatory scoring decisions.

For a high-stakes test like a specialist examination, valid pass–fail decision making demands a certain level of reliability.[27] The JPCA examination used in our study showed that no reliability of each component score exceeded 0.90, and only Portfolio was acceptable for summative purpose with a reliability score between 0.85 and 0.9. Reliability of noncompensatory scores (equal to current conjunctive standard) for CSA and MCQ was below 0.70, a level not allowed for pass/fail decision making. However, it is acceptable if we use a compensatory standard. Changing from noncompensatory to compensatory pass–fail decisions should be considered, yet whether JPCA's leaders accept the change or not is another issue because such a change in the rule could influence the mind-set of future candidates.

This study is based on a single board certification examination, with specialty-specific assessment content and learner sample. However, our data come from a national testing population and include data across three consecutive years, which should increase the generalizability of our findings. Moreover, our data and simulations are based on expert-defined weights, which may differ depending on the assessment context. Yet, we hope that findings from this study could shed light on the overall significance and add to the discussion on deeper understanding of compensatory and noncompensatory pass–fail decision making.

Assessments in the CBME era should consider balancing assessment tools measuring distinct but related competencies. We conclude that educators should investigate the impact of noncompensatory scoring by examining its measurement characteristics, in addition to curricular, clinical, and patient safety considerations.

Funding/Support: None reported.

Other disclosures: None reported.

Ethical approval: This study was approved by the institutional review board at the University of Illinois at Chicago.

H. Onishi is assistant professor, International Research Center for Medical Education, Graduate School of Medicine, The University of Tokyo, and vice chair, Committee for Specialist Certification, Japan Primary Care Association, Tokyo, Japan; ORCID: http://orcid.org/0000-0002-6979-1088.

Y.S. Park is associate professor, Department of Medical Education, University of Illinois at Chicago College of Medicine, Chicago, Illinois; ORCID: http://orcid.org/0000-0001-8583-4335.

R. Takayanagi is director, Gumma Family Medicine Center, Maebashi Kyoritsu Clinic, and vice chair, Committee for Specialist Certification, Japan Primary Care Association, Tokyo, Japan.

Y. Fujinuma is director, Centre for Family Medicine Development, Japanese Health and Welfare Co-operative Federation, and chair, Committee for Specialist Certification, Japan Primary Care Association, Tokyo, Japan.

References

1 Corcoran J, Downing SM, Tekian A, DaRosa DA. Composite score validity in clerkship grading. Acad Med. 2009;84(10 suppl):S120–S123.
2 Hicks PJ, Margolis M, Poynter SE, et al; APPD LEARN-NBME Pediatrics Milestones Assessment Group. The Pediatrics Milestones Assessment Pilot: Development of workplace-based assessment content, instruments, and processes. Acad Med. 2016;91:701–709.
3 Schwartz A, Margolis MJ, Multerer S, Haftel HM, Schumacher DJ; APPD LEARN–NBME Pediatrics Milestones Assessment Group. A multi-source feedback tool for measuring a subset of pediatrics milestones. Med Teach. 2016;38:995–1002.

4 Park YS, Lineberry M, Hyderi A, Bordage G, Xing K, Yudkowsky R. Differential weighting for subcomponent measures of integrated clinical encounter scores based on the USMLE Step 2 CS examination: Effects on composite score reliability and pass–fail decisions. Acad Med. 2016;91(11 suppl):S24–S30.
5 Margolis MJ, Clauser BE, Swanson DB, Boulet JR. Analysis of the relationship between score components on a standardized patient clinical skills examination. Acad Med. 2003;78(10 suppl):S68–S71.
6 Harik P, Clauser BE, Grabovsky I, Margolis MJ, Dillon GF, Boulet JR. Relationships among subcomponents of the USMLE Step 2 Clinical Skills examination, the Step 1, and the Step 2 Clinical Knowledge examinations. Acad Med. 2006;81(10 suppl):S21–S24.
7 Clauser BE, Balog K, Harik P, Mee J, Kahraman N. A multivariate generalizability analysis of history-taking and physical examination scores from the USMLE Step 2 Clinical Skills examination. Acad Med. 2009;84(10 suppl):S86–S89.
8 Baldwin P. Weighting components of a composite score using naïve expert judgments about their relative importance. Appl Psychol Meas. 2015;39:539–550.
9 Feldt LS. Estimating the reliability of a test battery composite or a test score based on weighted item scoring. Meas Eval Couns Dev. 2004;37:184–189.
10 Kane M, Case SM. The reliability and validity of weighted composite scores. Appl Meas Educ. 2004;17:221–240.
11 Messick S. Standards of validity and the validity of standards in performance assessment. Educ Meas. 1995;14:5–8.
12 Federation of the State Medical Boards and the National Board of Medical Examiners. 2017 Step 2 Clinical Skills (CS) content description and general information. http://www.usmle.org/pdfs/step-2-cs/cs-info-manual.pdf. Accessed July 20, 2018.
13 American Board of Surgery. Training and certification: General surgery. http://www.absurgery.org/default.jsp?examoffered_gs. Published 2017. Accessed July 20, 2018.
14 Association of American Medical Colleges. Core entrustable professional activities for entering residency: Curriculum developers' guide. https://www.mededportal.org/icollaborative/resource/887. Accessed July 20, 2018.
15 Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system—Rationale and benefits. N Engl J Med. 2012;366:1051–1056.
16 Yudkowsky R, Tumuluru S, Casey P, Herlich N, Ledonne C. A patient safety approach to setting pass/fail standards for basic procedural skills checklists. Simul Healthc. 2014;9:277–282.
17 Clauser BE, Harik P, Margolis MJ. A multivariate generalizability analysis of data from a performance assessment of physicians' clinical skills. J Educ Meas. 2006;43:163–191.
18 Swygert KA, Balog KP, Jobe A. The impact of repeat information on examinee performance for a large-scale standardized-patient examination. Acad Med. 2010;85:1506–1510.
19 Norcini JJ, Swanson DB, Grosso LJ, Webster GD. A comparison of several methods for scoring patient management problems. In: Kerby S, compiler. Proceedings of the Twenty-Second Annual Conference on Research in Medical Education. Washington, DC: Association of American Medical Colleges; 1983.
20 Webster GD, Shea JA, Norcini JJ, Grosso LJ, Swanson DB. Strategies in comparison of methods for scoring patient management problems: Use of external criteria to validate scores. Eval Health Prof. 1988;11:231–248.
21 Brennan RL. Generalizability Theory. New York, NY: Springer-Verlag; 2001.
22 Webb NM, Shavelson RJ, Haertel EH. Reliability coefficients and generalizability theory. In: Rao CR, Sinharay S, eds. Handbook of Statistics. Amsterdam, the Netherlands: Elsevier; 2006:81–124.
23 Livingston SA, Lewis C. Estimating the consistency and accuracy of classification based on test scores. J Educ Meas. 1995;32:179–197.
24 Park YS, Hyderi A, Bordage G, Xing K, Yudkowsky R. Inter-rater reliability and generalizability of patient note scores using a scoring rubric based on the USMLE Step-2 CS format. Adv in Health Sci Educ. doi: 10.1007/s10459-015-9664-3
25 Fleiss JL. Measuring nominal scale agreement among many raters. Psych Bull. 1971;76:378–382.
26 Gruppen L, Frank JR, Lockyer J, et al; ICBME Collaborators. Toward a research agenda for competency-based medical education. Med Teach. 2017;39:623–630.
27 Hambleton RK, Slater SC. Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. Appl Meas Educ 1997;10:19–38.