PUBLIC RELEASE

Technical Report: March 2018 CKE 2

Human Resources Professionals Association

15 May 2018

ICKETT
MEASUREMENT SYSTEMS

Contents

List of Tables

List of Figures

Executive Summary[1]

Note that this technical report covers only the primary new form or forms administered during an administration, and not detailed results for all forms used (which may include previously used forms, scrambled forms, and other modifications to maintain exam and score integrity).

The Comprehensive Knowledge Exam 2 (CKE 2) was administered to 181 candidates using computer-based testing at Prometric test centres from March 5 to 19, 2018, inclusive. The examination comprised 250 four-option multiple choice items and had a 5-hour time limit.

As per the CKE 2 blueprint, the exam was scored using the 225 best performing items (while adhering to the prescribed distribution across functional areas). The mean score for first-time candidates (n=143[2]) was 150.7 (67.0%), and for all candidates it was 145.3 (64.6%). Reliability was strong at .93. The final set of scored items adhered to the blueprint parameters.

The pass mark was set using equating back to the October 2017 and June 20176 CKE 2 administrations, yielding an integer pass mark of 135. Equating was conducted to compensate for minor changes in exam form difficulty so that any given candidate has an equivalent hurdle regardless of when they write the CKE 2. This pass mark resulted in a pass rate for first-time candidates of 72.7% and a pass rate for all candidates of 64.1%.

This report, the analyses performed, and the processes followed are consistent with NCCA standards[3] and ISO 17024 standards.[4]

[1] This technical report is an abbreviated version of the full report. Information has been excluded that if known to candidates could negatively affect the validity of future candidate test score interpretations. This includes item-level statistics, some information about the construction of test forms, and some specific details concerning equating.

[2] Excludes those who had failed an HRPA examination in the past or who were identified as being statistical outliers, and excludes those who wrote a form other than a primary form.

[3] National Commission for Certifying Agencies (2014). *Standards for the accreditation of certification programs*. Washington, DC: Institute for Credentialing Excellence.

[4] International Organization for Standardization (2012). *ISO/IEC 17024:2012 Conformity assessment – General requirements for bodies operating certification of persons*. Geneva: International Organization for Standardization.

Administration

Form Setting

Using only validated test items, Wickett Measurement Systems prepared one 250-item test forms (using a combination of scored and experimental test items). Wickett selected the final test forms according to the following parameters:

1. Including only items validated by the validation panel in the past two years
2. Fitting the total item count of 250
3. Absence of enemy items
4. Hitting the blueprint target value (+/− 3%) for each functional area
5. Maximizing spread across competencies
6. Reducing item exposure
7. Perceived psychometric effectiveness of the item, using statistics from previous administrations as available

The forms were proofed by Wickett for text errors and detection of potential enemy items. The form was presented to two CHRL EVC members (Kristin Rivait and Nancy Richard) for review item currency and further detection of enemy items on January 4, 2018. As per their input, seven items were replaced and the form finalized.

The final form composition for the primary March 2018 CKE 2 form is shown in Table 1. All functional areas are within two items of their targets, and as such, the two forms reflect the blueprint.

Note that at any administration, HRPA makes use of previously validated and administered test forms along with new test forms, in addition to employing other mechanisms to maintain the integrity of the exams and candidates scores.

Table 1: Test form as administered

	Functional Area	Actual Items	Target	Variance
10	Strategy	25	25	—
20	Professional Practice	30	30	—
30	Organizational Effectiveness	35	35	—
40	Workforce Planning & Talent Management	35	35	—
50	Labour & Employee Relations	25	25	—
60	Total Rewards	25	25	—
70	Learning & Development	25	25	—
80	Health, Wellness & Safe Workplace	20	20	—
90	HR Metrics, Reporting & Financial Management	30	30	—
	TOTAL	**250**	**250**	**—**

Testing Window

The examination was administered via computer-based testing at Prometric test sites primarily in Ontario. The testing window was from March 5 to 19, 2018, inclusive, and 181 candidates wrote the exam.

Candidates had access to a basic-function calculator on screen. No other aids or sources were allowed.

Analysis

Data Cleaning and Integrity Checks

Prometric provided data in .xml format via a secure ftp site. Candidate files were provided as candidates completed the examination throughout the testing window. These files were extracted to Microsoft Excel for processing. They contained identifying information for each candidate, form information, start and stop times, answer string, key string, candidate total score, item comments if the candidate made any, and time spent per item.

The data files received were reconciled against the roster provided by Prometric to ensure that all .xml files had been received. Further, the candidate total score as computed by Prometric was reconciled with that computed by Wickett for the full set of 250 items to verify key accuracy. Comments on items were also reviewed to identify any specific item-level issues. No item problems were uncovered.

The average time taken by all candidates was assessed to detect potential examination timing concerns. The distribution is shown in Figure 1. The mean was 3 hours, 39 minutes (7 minutes than in October 2017). The time limit on the CKE 2 was 5 hours, suggesting that time was not a factor in scores across candidates.

Eight (4%) of candidates took the full 5 hours suggesting that those candidates may have wanted more time, and 8 candidates (4%) left at least one item blank suggesting those candidates timed out of the exam before being able to complete it. These metrics will continue to be monitored, and at the present do not appear problematically high.

The correlation between scores on the 250 items and time spent writing the examination was negligible at a value of .08, suggesting that time constraints were not generally an issue for candidate performance. (Note that two candidates exceeded the time limit; these candidate were granted additional time in advance of the administration as an accommodation.)

Candidate scores were computed across the window to look for any evidence of item exposure. As shown in Figure 2 there was little variation across the window. The difference between the first three days and the last three days was a negligible increase of 0.9 marks out of 250.

As a matter of interest, candidate volumes were also examined across the window and these are also shown in Figure 2. Oddly, the usual pattern of increased sittings at the end of the window was not seen at this administration. This has not been explained but is not considered psychometrically relevant.

Figure 1: Examination time distribution for all candidates

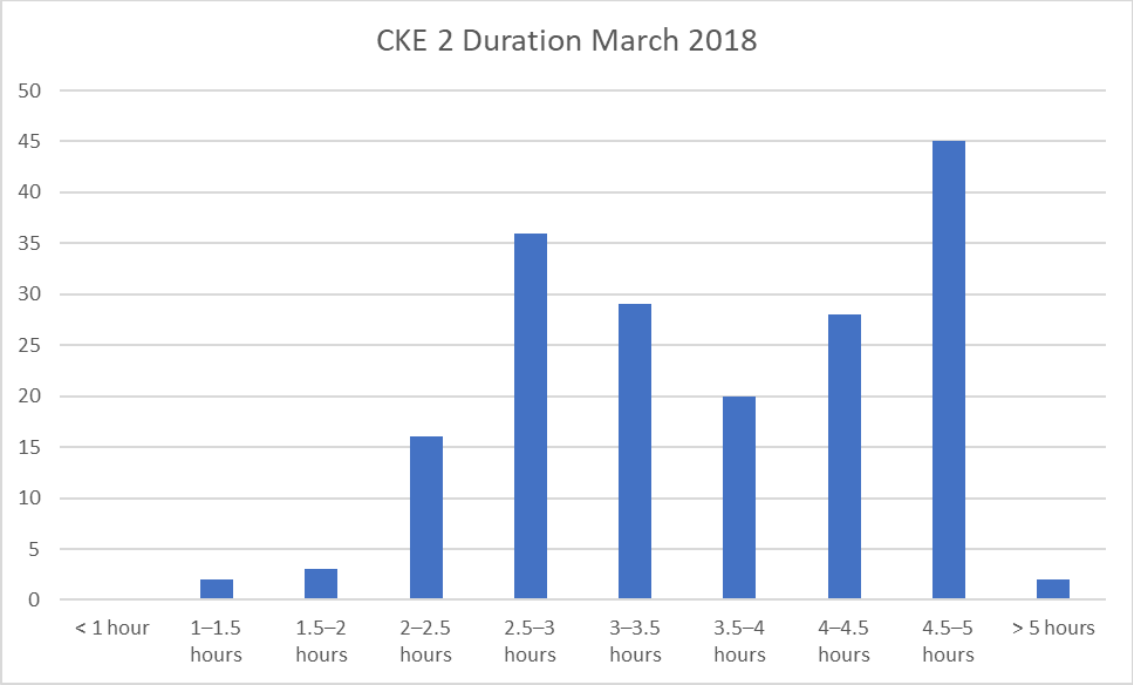
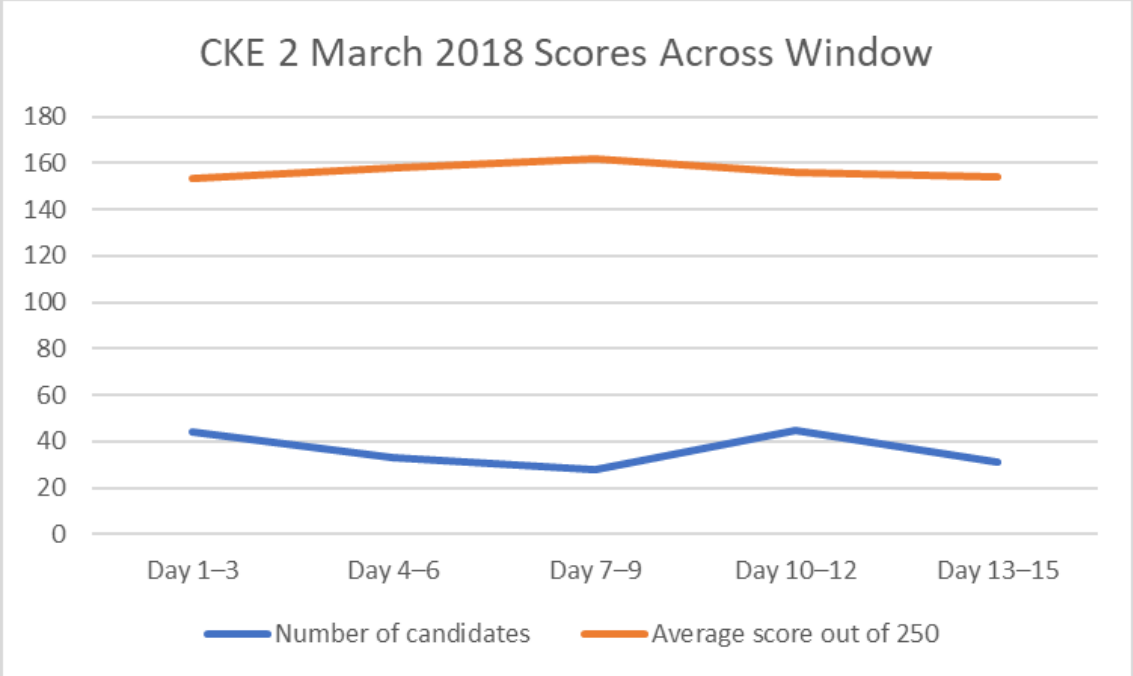


Figure 2: Candidate volume and score trends across testing window



After removing candidates who were administered a previously-used test form (who were scored using the same decisions employed at the time that form was originally used), scores were created for all remaining candidates based on the full set of 250 items. No candidates were flagged for an abnormally low or high score (*z* value outside +/− 3.0). Also, the 250 items

were arbitrarily broken into blocks of 25 items for each candidate; the 10 resulting scores for each candidate were evaluated for outliers as well. For candidates with any subscore more than 3 standard deviations (SD) from their average *z*-score, the .xml file was examined closely for any issues. All outliers were removed from initial analyses and candidates with abnormal response patterns were also removed. To be conservative, candidates who had been granted a testing accommodation were also removed from the main analysis (simply because their testing conditions were not the same as the main group of candidates, and even though each accommodation was granted on the premise that it would make the testing experience equivalent in terms of opportunity to demonstrate competence). As a result of all of these factors, two candidates were removed from analysis.

Candidates who had failed a previous HRPA examination (CKE, CKE 1, or CKE 2) scored lower than did those who had not (54.7% and 64.4%, respectively, on the full exam of 250 items). This difference was meaningful and significant ($t(74)=6.56$, $p<.001$); in keeping with standard procedures, these candidates were removed from subsequent analyses. The CKE 2 analysis proceeded with 143 candidates.

Owing to the modest number of candidates, all subsequent analyses were interpreted with caution.

Post-Examination Survey

Candidates were provided access to the post-examination survey immediately after submitting their responses to the CKE 2; 176 responses were obtained from candidates (response rate, 97%).

Table 2 shows the responses to the administration-related questions for CKE 2 candidates. Note that candidates were generally very positive about the administration experience. One exception to this is Question 1 on booking at a convenient time which showed up with lower endorsement than usual; this may be related to candidates being more evenly dispersed across the time window. Table 3 shows the content-related questions for CKE 2 candidates. There was a tendency to more neutrality on these questions. The lower rating for perceived fairness (Question 14) warrants monitoring as it continues to be low.

Candidates were asked to express their opinion regarding whether completing the examination on a computer affected their performance. Table 4 shows that most candidates felt it made no difference, and that where a preference was expressed it was essentially equally split between those who preferred computer and those who did not.

An open-ended question was also posed to candidates asking for any additional comments. Those comments were provided to HRPA for information and consideration. Nothing in the comments or survey data raised concerns about item analysis or scoring.

Table 2: Administration-related post-examination survey questions*

	Question	SA	A	N	D	SD	Score	Agreement
1.	I was able to book a seat to write the examination at a time that was convenient for me.	50	53	10	39	24	3.4	59%
2.	I was well informed about what documents to bring to the exam location.	112	54	4	3	1	4.6	95%
3.	Proctors enforced the exam-day rules and the security procedures at the test centre were what I expected.	114	53	3	3	1	4.6	96%
4.	Proctors were professional and courteous.	120	49	4	1	0	4.7	97%
5.	The tutorial helped me understand how to complete the examination on the computer.	86	79	5	1	1	4.4	96%
6.	Navigation through the examination was easy and intuitive.	105	62	2	4	1	4.5	96%

*Response categories: SA = Strongly Agree; A = Agree; N = Neutral; D = Disagree; SD = Strongly Disagree.

Table 3: Content-related post-examination survey questions*

	Question	SA	A	N	D	SD	Score	Agreement
7.	The time allotted for this examination was sufficient.	105	57	3	8	1	4.5	93%
8.	Information available prior to exam day provided me with adequate details about the content and format of the exam.	29	67	30	34	14	3.4	55%
9.	I feel I was adequately prepared to write this examination.	8	62	55	37	12	3.1	40%
10.	The questions in the examination were clearly written.	11	69	49	40	6	3.2	46%
11.	The terminology used in the examination was accurate.	16	87	50	17	4	3.5	59%
12.	The situations presented in the examination were realistic.	16	119	29	8	2	3.8	78%
13.	The questions in the examination reflected the examination blueprint.	6	71	63	21	6	3.3	46%
14.	The examination was a fair assessment of my ability.	4	41	67	41	20	2.8	26%

*Response categories: SA = Strongly Agree; A = Agree; N = Neutral; D = Disagree; SD = Strongly Disagree.

Table 4: Preference regarding computer-based testing versus pencil-and-paper

Question	Count	%
I feel that completing the examination on a computer improved my performance.	25	14%
I feel that completing the examination on a computer decreased my performance.	31	18%
I feel that completing the examination on a computer had no effect on my performance.	118	68%

Initial Analysis

The full CKE 2 examination was 250 items, of which 225 were to be scored. The other 25 items were designated as experimental items. However, because only one new form was administered, all items were potentially available for scoring and the focus of subsequent item analysis and key validation was on determining the best set of 225 items that still reflected the examination blueprint.

The initial analysis summary statistics for the new form are presented in Table 5.

Table 5: Initial examination statistics

Index	CKE 2
Eligible items	250
Candidates	181
Candidates in analysis	143
Mean	161.7 (64.7%)
Range	99–223 (39.6–89.2%)
Cronbach's alpha	.93
Mean r_{pb}*	.21

Standard classical test theory analysis was conducted to identify the following:

1. Item difficulty (percent obtaining correct result, *p*)
2. Item discrimination (corrected point-biserials, r_{pb}*)
3. Distractor quality (based primarily on distractor discrimination)

Wickett compiled these statistics, along with any comments made by candidates concerning flagged items, to identify items that may have been keyed incorrectly or that were performing

poorly. Most emphasis was placed on the corrected point-biserials as evidence of item quality. Items were ranked from worst performing to best performing accordingly.

Key Validation

Key validation was conducted via web meeting on March 23, 2018, using the CHRL Exam Validation Committee (EVC). The EVC (Table 6) was first trained in basic item and test analysis methods and was oriented to the main statistics used to evaluate the quality of the CKE 2.

Table 6: Key validation panel members

Member	Credential	Years of Relevant Experience	Year on EVC	Industry
Jennifer Borges	CHRL	10–14	1	Manufacturing
Debbie Brandt	CHRL	10–14	1	Government and public centre agencies
Annette Dhanasar*	CHRL	15–19	1	Transportation
Christine Kelsey	CHRL	1–4	1	Entertainment
Jennifer King*	CHRL	20–29	1	Business and professional services
Nancy Richard*	CHRL	15–19	1	Regulation/Public sector
Kristin Rivait*	CHRL	15–19	1	Health care
Lisa Scian	CHRL	15–19	1	Information & communication technology

*Unable to participate.

The group was informed that test reliability, as measured by Cronbach's alpha, was .928 based on the set of 250 potentially scored items and that this was above the generally accepted threshold of .80. They were also informed that part of the goal of the key validation review was to bring this value up if possible.

The group was walked through the flagged items one at a time, with the recommendation that the worst performing items be removed from scoring but less direction on those items with borderline statistics. Where available, any comments made by candidates on the items were also shown. The group made the decision based on content and the data through discussion; they removed the 25 items that they felt were least appropriate to retain for scoring. Past item data were also used where available, and the group was directed consider it an addition to statistics from the June administration. Comments made by the panel members about specific items were recorded for future item revision activities.

Not all remaining items were strong-performing, and several items were retained that were very easy or very hard or that had a low corrected point-biserial. However, most were moderate to strong items. The final alpha for the set of 225 scored items was .931. The difficulties ranged from 28.7% to 93.7%, with a mean of 67.0%. The r_{pb}* values ranged from .02 to .51 with a mean of .23.

Table 7 presents the scored CKE 2's final fit to the examination blueprint. In all cases, the final number of scored items within a functional area fit within the established range.

The group endorsed the final set of items for use in scoring the March 2018 CKE 2 candidates who took this form.

Table 7: Final scored examination fit to blueprint

	Functional Area	Actual	Min.	Target	Max.	Blueprint Range
10	**Strategy**	23	17	22.5	28	10% ± 2.5%
20	**Professional Practice**	28	22	27	32	12% ± 2.5%
30	**Organizational Effectiveness**	30	26	31.5	37	14% ± 2.5%
40	**Workforce Planning & Talent Management**	31	26	31.5	37	14% ± 2.5%
50	**Labour & Employee Relations**	21	17	22.5	28	10% ± 2.5%
60	**Total Rewards**	23	17	22.5	28	10% ± 2.5%
70	**Learning & Development**	22	17	22.5	28	10% ± 2.5%
80	**Health, Wellness & Safe Workplace**	17	13	18	23	8% ± 2.5%
90	**HR Metrics, Reporting & Financial Management**	30	22	27	32	12% ± 2.5%
Total		**225**				

Establishing the Pass Mark: Equating

Equating, as per Kolen and Brennan (2014),[5] was used to establish the pass mark for the March 2018 CKE 2. The goal of this process was to set a pass mark for the March 2018 CKE 2 that would be equivalent to that set for previous CKE 2 administration; that is, to set a pass mark that would give each candidate the same probability of passing regardless of which form they took.

The passing standard for the CKE 2 was originally set after the November 2015 offering of the CKE 2 using the Modified Angoff method. General details on that method can be found in Appendix A. Specific information on the standard-setting session is provided in the Technical Report issued for the November 2015 administration.

[5] Kolen, M.J., & Brennan, R.L. (2014). *Test equating, scaling, and linking.* New York, NY: Springer.

To pass the CKE 2, a candidate must meet or surpass the overall test pass mark and meet or surpass the threshold set for each of the nine functional areas. These thresholds are set independently and are described in turn.

Total Score Pass Mark

Two equating procedures were conducted back to different administrations (June 2017 and October 2017). The intention following these two equating runs is to average them to arrive at a final pass mark for the March 2018 CKE 2. These administrations were chosen as the most recent administration and the administration corresponding roughly to the same administration month the previous year. The March 2017 administration was not used in this procedure because of the very low number of candidates sitting that administration.

Equating back to the October 2017 Administration

Linear equating was the chosen method for setting the pass mark. Linear equating is preferred with more than 100 candidates, and equipercentile equating is preferred with more than 1,000 candidates. With candidate samples of fewer than 100, mean or circle arc equating is most prudent.

All candidates in the analysis (i.e., no repeat candidates or outliers) were used in the equating process. Delta plot analysis was used to identify anchor items showing substantial deviations (generally, although not exclusively, greater than three SD units) from expected difficulty values, with an emphasis on establishing an anchor set with difficulty equivalent to that of the full form (and equivalent within each functional area) that adhered to the blueprint. Further, items with very high or low difficulty values and those with low corrected point-biserials were also flagged for potential removal from the anchor set. The goal was a strong midi-test (i.e., moderate range of difficulty, moderate to high discrimination, fit to blueprint) of sufficient length to estimate candidate ability.

The selected set of anchor items had a mean difficulty of 0.67 and a mean corrected point-biserial of .23 (for March 2018 candidates).

Table 8 shows the fit of the set of anchor items to the blueprint, as percentages. The actual counts are well-aligned with targets and reflect the scope and approximate weighting across the full exam.

Table 8: Anchor item fit to blueprint – to October 2017

Area*	Actual	Target
10	10%	10%
20	12%	12%
30	14%	14%
40	12%	14%
50	9%	10%
60	10%	10%
70	10%	10%
80	8%	8%
90	12%	12%

*See Table 7 for the full name of each functional area.

Mean, Tucker, and Levine observed-score methods were computed to ascertain concordance of solutions. Given the sample sizes and similarities of test parameters, Tucker equating was considered the preferred method.

Table 9 shows some of the parameters used to derive the equating estimates, along with other parameters describing the test forms. Of note is that on the anchor items, the population taking the March 2018 CKE 2 scored marginally less than the population taking the October 2017 CKE 2 (66.9% vs. 67.4%, respectively; $t(376)=0.43$, *ns*). Because the March 2018 CKE 2 candidates scored at about the same level (based on the anchors), they would likely have about the same pass rate (or slightly lower) as was seen in October.

The equating analysis bears this out (Table 10) for the most part. The Tucker, Levine observed, and mean methods indicate a pass mark of 135–136. The pass rate based on this equating run is somewhat lower than what was seen in October 2017. The Tucker equating value of 134.48 was extracted from this analysis for use in setting the final pass mark.

Table 9: Equating parameter table – Total pass mark, to October 2017

		2017	2018
		October	March
	n	235	143
	Scored items	225	225
Mean score	Total	66.0%	67.0%
Mean score	Anchors	67.4%	66.9%

Table 10: Equating outcome table – Total pass mark, to October 2017

	Pass Mark		Pass Rate	
Method	**Precise**	**Integer**	**All**	**First Time**
Combo October 2017	**132.37**	**133**	**69.3%**	**75.7%**
Tucker	134.48	135	*64.1%*	*72.7%*
Levine observed	134.54	135	*64.1%*	*72.7%*
Mean	135.70	136	*62.4%*	*71.3%*

Equating back to the June 2017 Administration

Linear equating was the chosen method for setting the pass mark, given the sample sizes involved.

All candidates in the analysis (i.e., no repeat candidates or outliers) were used in the equating process. Delta plot analysis was used to identify anchor items showing substantial deviations (generally, although not exclusively, greater than three SD units) from expected difficulty values, with an emphasis on establishing an anchor set with difficulty equivalent to that of the full form (and equivalent within each functional area) that adhered to the blueprint. Further, items with very high or low difficulty values and those with low corrected point-biserials were also flagged for potential removal from the anchor set. The goal was a strong midi-test (i.e., moderate range of difficulty, moderate to high discrimination, fit to blueprint) of sufficient length to estimate candidate ability.

The selected set of anchor items had a mean difficulty of 0.67 and a mean corrected point-biserial of .25 (for March 2018 candidates).

Table 11 shows the fit of the set of anchor items to the blueprint, as percentages. The actual counts are well-aligned with targets and reflect the scope and approximate weighting across the full exam.

Table 11: Anchor item fit to blueprint – to June 2017

Area*	Actual	Target
10	11%	10%
20	12%	12%
30	13%	14%
40	12%	14%
50	11%	10%
60	11%	10%
70	11%	10%
80	9%	8%
90	13%	12%

*See Table 7 for the full name of each functional area.

Mean, Tucker, and Levine observed-score methods were computed to ascertain concordance of solutions. Given the sample sizes and similarities of test parameters, Tucker equating was considered the preferred method.

Table 12 shows some of the parameters used to derive the equating estimates, along with other parameters describing the test forms. Of note is that on the anchor items, the population taking the March 2018 CKE 2 scored marginally lower than the population taking the June 2017 CKE 2 (66.6% vs. 67.4%, respectively; $t(340)=0.57$, *ns*). Because the March 2018 CKE 2 candidates scored at about the same level (based on the anchors), they would likely have about the same pass rate (or slightly lower) as was seen in June.

The equating analysis bears this out (Table 13). The Tucker, Levine observed, and mean methods indicate a pass mark of 136–137. The pass rate based on this equating run is somewhat lower than in June, as expected. The Tucker equating value of 135.06 was extracted from this analysis for use in setting the final pass mark.

Table 12: Equating parameter table – Total pass mark, to June 2017

		2017	2018
		June	March
	n	199	143
	Scored items	225	225
Mean score	Total	65.5%	67.0%
	Anchors	67.4%	66.6%

Table 13: Equating outcome table – Total pass mark, to June 2017

	Pass Mark		Pass Rate	
Method	**Precise**	**Integer**	**All**	**First Time**
Combo June 2017	**131.06**	**132**	**69.3%**	**75.7%**
Tucker	135.06	136	*62.4%*	*71.3%*
Levine observed	135.10	136	*62.4%*	*71.3%*
Mean	136.00	137	*61.3%*	*69.9%*

Combined Results

Table 14 shows the pass mark values across the three equating runs. The value in green highlight is the one that would be selected based on population parameters at each equating run. Though different weighting and averaging methods were considered, all resulted in a value between 134 and 135. Without a clear reason to do otherwise, the simple arithmetic mean (134.771878) of the two identified values was the recommended pass mark for the March 2018 CKE 2.

Using the established convention for this testing program, the mean value was rounded up to a cut score of 135. The resulting pass rate of 72.7% for first-time candidates is somewhat lower than seen in June and October 2017 (see Table 15), but there was evidence the March candidates were slightly lower in ability and so this was consistent. The pass rate for all candidates was 64.1%.

Table 14: Equating outcome table – Combined results, Total pass mark

	Oct17	Jun17
Tucker	134.5	135.1
Levine observed	134.5	135.1
Mean	135.7	136.0

Table 15: Historic pass rates – Total pass mark

	Pass rate	
	All	1st time
Jun16	73.8%	78.5%
Nov16	64.6%	68.1%
Mar17	59.4%	72.9%
Jun17	70.0%	75.9%
Oct17	69.3%	75.7%
Mar18	**64.1%**	**72.7%**

Functional Area Minimum Thresholds

The original functional area minimum thresholds were established in November 2015 to identify candidates who scored egregiously low on any individual functional area (see Appendix B for a conference presentation made regarding this method). Since that time, equating has been employed to produce equivalent thresholds on subsequent administrations.

Tucker equating was employed for each functional area when equating back to October and June 2017, as this was the method selected for the total test score equating in those equating runs. The decisions outlined above to finalize anchor selection for the total test score equating were made so that they would also be appropriate to equating at the functional area level.

Table 16 shows alignment between anchor performance and full exam functional area score. The goal of close alignment was generally achieved though there were exceptions. By equating across two administrations, any existing biasing effects should be reduced.

The resulting thresholds across each equating run are shown in Table 17.

Table 18 shows the outcomes and other relevant information related to equating of functional area thresholds. Note that two candidates failed the exam based solely on having missed the threshold on a functional area.

Table 19 shows the outcomes for each decision criterion. About one-half of the failing candidates failed both at the total score level and at the functional area level, whereas the remainder failed based only on the total score pass mark.

The thresholds for all functional areas, and the process used to derive them, were presented to the same panel used for key validation (Table 6) via teleconference on March 29, 2018. The panel approved the functional area thresholds (which were presented along with the consequent pass rate consequences) for recommendation to HRPA. HRPA subsequently accepted the recommendation from the panel and the functional area thresholds were formally established.

Table 16: Alignment between difficulty of anchors and full exam

Area*	October 2017 Anchors	June 2017 Anchors	Full Exam
10	63%	66%	63%
20	74%	73%	73%
30	66%	66%	68%
40	69%	73%	67%
50	68%	65%	67%
60	68%	63%	68%
70	63%	63%	66%
80	73%	70%	71%
90	60%	61%	59%

*See Table 7 for the full name of each functional area.

Table 17: Equating outcome table – Combined results, Functional area thresholds

		10	20	30	40	50	60	70	80	90
To Oct17	Tucker	8.43	11.98	11.09	10.74	8.05	8.81	8.10	6.60	8.53
To Jun17	Tucker	8.34	11.78	10.80	10.26	8.30	8.97	8.23	6.67	8.58
	Average	8.39	11.88	10.94	10.50	8.17	8.89	8.16	6.63	8.55
	Integer	9	12	11	11	9	9	9	7	9

Table 18: Equating summary table – Functional area thresholds

Area*	Cut[i]	Integer[ii]	Items	Cut as %	Previous Cut %[iii]	Alpha[iv]	Mean	Unique Fails[v]
10	**8.39**	9	23	36%	37%	.63	14.6	**0**
20	**11.88**	12	28	42%	38%	.65	20.6	**0**
30	**10.94**	11	30	36%	38%	.71	20.5	**0**
40	**10.50**	11	31	34%	37%	.66	20.8	**0**
50	**8.17**	9	21	39%	39%	.56	14.1	**0**
60	**8.89**	9	23	39%	36%	.54	15.7	**0**
70	**8.16**	9	22	37%	40%	.62	14.5	**1**
80	**6.63**	7	17	39%	29%	.40	12.1	**1**
90	**8.55**	9	30	29%	31%	.66	17.8	**0**

*See Table 7 for the full name of each functional area.

[i]Threshold set through equating.

[ii]Rounded up value of cut score as used for making candidate decisions.

[iii]Threshold set on previous administration.

[iv]Cronbach's alpha for functional area.

[v]Number of candidates failing based on not meeting the functional area threshold who otherwise passed at the total score level.

Table 19: Passing decisions – Total pass mark and functional areas

Fails	Both measures	31	17.1%
	Total score only	34	18.8%
	Functional area score only	2	1.1%
Passes	Neither	114	63.0%

Scoring

To finalize the scoring, repeat and outlier candidates who were not included in the item and form analysis were reinserted into the dataset. Scores for each of the nine functional areas were also computed for each candidate. An Excel file with the final candidate results was provided to HRPA.

Table 20 provides the means and standard deviations for the functional areas and for the total score, using all candidates who took the new March 2018 CKE 2 form. Table 21 provides the correlations between all functional areas. Caution should be exercised in interpreting differences between correlations. Variation can be explained largely by the number of items making up each functional area score. That is, functional areas with fewer items on the exam have lower correlations with the other functional areas. Figure 3 shows the distribution of scores for all candidates, along with the pass mark.

Table 20: Total and functional area scores for all candidates

	Functional Area	Percentage	Mean	SD*
10	**Strategy**	61%	14.1	3.4
20	**Professional Practice**	71%	20.0	3.6
30	**Organizational Effectiveness**	65%	19.6	4.6
40	**Workforce Planning & Talent Management**	65%	20.0	4.3
50	**Labour & Employee Relations**	65%	13.6	3.0
60	**Total Rewards**	66%	15.1	3.3
70	**Learning & Development**	63%	14.0	3.5
80	**Health, Wellness & Safe Workplace**	70%	11.8	2.2
90	**HR Metrics, Reporting & Financial Management**	57%	17.2	4.2
Total score		64.6%	145.3	25.5

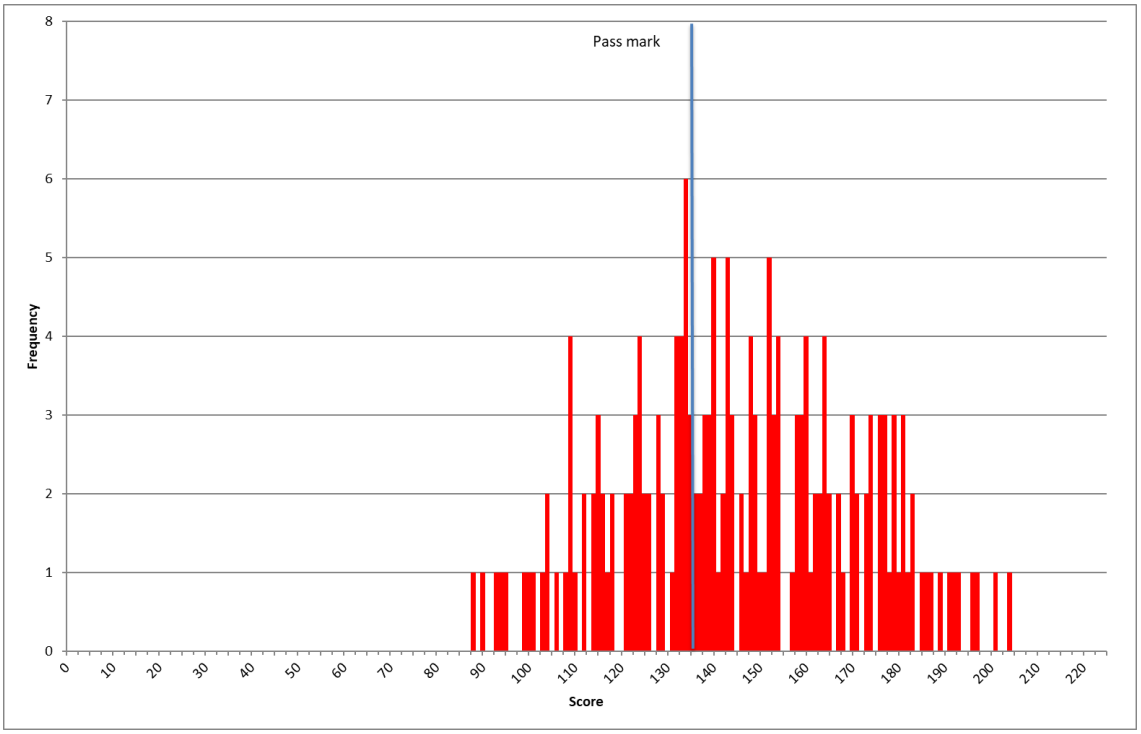
*SD = Standard deviation.

Table 21: Correlations between functional area scores for all candidates

Area*	10	20	30	40	50	60	70	80	90
10		.58	.61	.61	.58	.54	.55	.42	.56
20			.69	.70	.57	.55	.57	.43	.60
30				.71	.61	.67	.62	.51	.66
40					.67	.63	.62	.45	.63
50						.59	.56	.50	.60
60							.56	.46	.50
70								.55	.54
80									.49
90									

*See Table 20 for the full name of each functional area.

Figure 3: Score distribution for all candidates



Key Examination Metrics

Table 22 shows the key examination metrics for candidates included in the main analysis; that is, only first-time candidates, with outliers removed. Past metrics are provided for reference.

Note that March 2017 was the first computer-based testing delivery of the CKE 2, but this was not considered material to data analysis or interpretation.

Table 22: Key examination metrics – Candidates included in analysis only

Index	March 2018	October 2017	June 2017	March 2017	November 2016
Scored items	225	225	225	225	225
Candidates	143	235	199	70	339
Mean	150.7 (67.0%)	148.5 (66.0%)	147.3 (65.5%)	151.3 (67.2%)	144.4 (64.2%)
Median	152 (67.6%)	150 (66.7%)	150 (66.7%)	156 (69.3%)	146 (64.9%)
Skewness	−0.132	−0.403	−0.295	−0.498	−0.112
Kurtosis[i]	−0.555	0.108	−0.306	−0.692	−0.503
Range	88–204 (39.1–90.7%)	78–203 (34.7–90.2%)	78–195 (34.7–86.7%)	93–188 (41.3–83.6%)	80–203 (35.6–90.2%)
Standard deviation	24.53	22.98	23.67	23.91	25.04
Cronbach's alpha	.93	.92	.93	.93	.93
Mean r_{pb}*	.23	.21	.23	.23	.23
SEM[ii]	6.43	6.52	6.34	6.34	6.50
SEM at the pass mark	6.89	6.94	6.74	6.79	6.87
Decision consistency (uncorrected)[iii]	.89	.90	.90	.92	.90
Perceived fairness[iv]	27%	28%	27%	31%	n/a
Pass mark	134.772	132.371	131.061	134.664	131.936
Effective pass mark	135	133	132	135	132
Pass rate	72.7%	75.3%	75.4%	72.9%	68.1%

[i]Excess

[ii]SEM = standard error of measurement.

[iii]Subkoviac method.

[iv]Based on responses to the post-examination survey. Value here differs from that presented in main body of report because this value includes only candidates in the analysis.

Related Development Activities

Since the last administration of the CKE 2 in October 2017, the following exam development activities have taken place.

Item Writing

To fill gaps in the bank and renew content, item writing was conducted in October–December 2017. Item writers (see Table 23) were identified by HRPA and trained in a remote session by Wickett on October 31, 2017.

Table 23: Item writers

Writer	Credentials	Years of Relevant Experience	Industry
Nita Chhinzer	PhD	15+	Education – Professor at Guelph University
Aaron Gordon	PhD	15+	Education – Professor at Algoma University
Gail Lawrence	CHRL	20+	Education/Consultant – Professor at Lakehead University

The item writers were provided with training via teleconference, and received printable files covering the main elements of the training. The general guidance for writing quality multiple choice items was drawn primarily from Haladyna & Rodriguez (2013).[6]

Each item writer was selected based on expertise in identified functional areas, and they were assigned items within those functional areas. More specifically, each item writer was assigned competencies (drawn from the *HRPA Professional Competency Framework* [2014]) that were to be the focus of their items. Item writers were assigned 30 items each to write, for a total of 90 items (one item writer completed 27 items, and so 87 new items were obtained).

The item writers had access to the style guide that governs language usage on the HRPA exams and were provided with recent text books as necessary. Item writers were required to include at least one authoritative source to back up each test item, and also provide rationales for the correct and incorrect answers.

Each item writer worked remotely, sending items to Wickett for review and comment via a secure file share site. Items were exchanged until such time as the item writer was comfortable with the content and Wickett was comfortable that the item would be successful at review, validation and upon use with candidates. This generally required several iterations per item.

[6] Haladyna, T. M., & Rodriguez, M.C. (2013). *Developing and validating test items.* New York, NY: Routledge.

Once all items were drafted and declared complete, they were sent a certified professional editor for editorial. Items were adjusted based on this input and comments noted if future reviewers would need to attend to specific content concerns.

Item Review

Following the item writing exercise in October–December 2017 there was need for group review of those items before moving them to formal validation and use on the CKE 2. The group had 133 items for consideration (taken predominately from the newly written items, supplemented with other unreviewed items in the bank required to fill gaps in the bank).

The review session was held February 12–14, 2018 at HRPA offices. The panel members who participated are shown in Table 24. This session involved the review of ELE items as well.

Table 24: Panel for item review session

Reviewer	Credentials	Years of Relevant Experience	Industry
Gabriella Fermo	CHRL	15–19	Distribution
Julie Jamieson	CHRL	20–29	Retail/Distribution
Vanessa Lewerentz	CHRL	20–29	Banking/Finance
Lynn Rivard	CHRL	10–14	Non-profit organization
Irene Stretton	CHRL	10–14	Industrial Distributor
Laurie Torno	CHRL	20–29	Educational Services

The panel members received training on the review activity, and then worked primarily individually reviewing items to make sure they reflected current practice. Where panel members proposed changes, these were discussed by the group before implementation.

The panel members reviewed and approved 110 items as suitable for CKE 2, moved 2 items to the CKE 1 bank, and rejected 21 items. Of the approved items, 14 saw text changes to the stem and/or options before approval.

The items were updated in the bank, and those that were approved were deemed ready for validation before use on future examinations.

Validation

To validate newly reviewed items and renew the validation of items expiring from usability, a validation session was held with the EVC March 12–14, 2018. The group had 250 items for consideration, and these items were to supplement the bank in support of the upcoming

administrations. The committee members who participated are shown in Table 25. This session involved the validation of ELE items as well.

Table 25: CHRL EVC members at validation

Member	Credential	Years of Relevant Experience	Year on EVC	Industry
Jennifer Borges*	CHRL	10–14	1	Manufacturing
Debbie Brandt	CHRL	10–14	1	Government and public centre agencies
Annette Dhanasar	CHRL	15–19	1	Transportation
Christine Kelsey*	CHRL	1–4	1	Entertainment
Jennifer King	CHRL	20–29	1	Business and professional services
Nancy Richard	CHRL	15–19	1	Regulation/Public sector
Kristin Rivait	CHRL	15–19	1	Health care
Lisa Scian*	CHRL	15–19	1	Information & communication technology

*Unable to participate.

The EVC members received advance materials outlining:

- Purpose of the session
- Description of the CHRL credential
- CKE 2 blueprint
- Criteria for good test items
- Validation process

The committee members received training on basic psychometrics and the validation activity, and then worked primarily individually reviewing items to make sure they reflected current practice and were suitable to make decisions about who should receive the CHRL credential. Where committee members proposed changes, these were discussed by the group before implementation.

For each item, the committee was asked to either:

- Validate the item for use in the next two years to make decisions about who would be certified as CHRL
- Move the item to the CKE 1 or ELE bank
- Revise the item to make it suitable for use

- Declare the item unsound and send it back for revision or removal from the bank

The bulk of the session saw the committee members reviewing items independently and submitting their assessments in blocks of approximately 20–25 items. Those assessments were tabulated and any items that were not validated as is by the full committee were discussed until there was agreement on changes and the future use of the item.

The committee reviewed and validated 245 items as suitable for CKE 2, moved 1 item to the ELE bank, and rejected 4 items. Very few items were revised as the items had gone through considerable review before getting to the committee for validation.

CKE 2 Blueprint Revision

At the validation session held March 12–14, 2018 (see section above), the CHRL EVC was asked to consider minor revisions to the CKE 2 blueprint.

The following changes were approved:

1. Formalization of purpose statement for the CKE 2:

   > "The CKE 2 assesses whether a candidate has the knowledge required to be an effective human resources professional at the CHRL level, in Ontario. Knowledge related exclusively to employment-related legislation will be assessed on the CHRL Employment Law Examination."

2. Change in how item counts were reported from "225 items plus 25 experimental items" to "250 items, of which 20–30 are experimental". This change was to allow more flexibility in form design.
3. Revision of weights for each functional area based on proportion of that functional area devoted to employment law. The CHRL EVC reviewed the *HRPA Professional Competency Framework* (2014) and identified competencies that should be assessed on the CHRL Employment Law Examination (which was created after the current CKE 2 blueprint was approved in 2015). Some competencies were fully removed from the CKE 2 blueprint, and some were allocated a 'half-weight' meaning that they should only be sampled at half the frequency of other competencies. As a result of this, some functional areas had their relative weight on the full exam increased (if they had no competencies allocated fully or partially to the ELE 2) and some had their weights decreased (if they had competencies allocated fully or partially to the ELE 2).

The final weights approved for the CKE 2 are shown in Table 26.

Table 26: Blueprint weights and ranges as approved by the CHRL EVC

		CKE 2	
Functional Area		**Weight**	Range
10	Strategy	11%	+/- 2%
20	Professional Practice	11%	+/- 2%
30	Organizational Effectiveness	14%	+/- 2%
40	Workforce Planning & Talent Management	14%	+/- 2%
50	Labour & Employee Relations	9%	+/- 2%
60	Total Rewards	10%	+/- 2%
70	Learning & Development	11%	+/- 2%
80	Health, Wellness, & Safe Workplace	8%	+/- 2%
90	Human Resources Metrics, Reporting, & Financial Management	12%	+/- 2%

The following competencies were declared only applicable to the CHRL ELE:

- C035
- C036
- C037
- C117
- C139
- C177
- C179
- C204
- C205

The following competencies received ‘half-weights,’ meaning that they would still be applicable to the CKE 2, but would appear only half as often as other competencies within that functional area:

- C038
- C118
- C119
- C131
- C133
- C134
- C182
- C187

These changes were considered to be minor in nature, and the relative weighting by content for the CKE 2 was not altered, excepting for content that was redirected to the CHRL ELE.

The changes were presented to the HRPA registrar on March 14, 2018 and were approved as recommended by the CHRL EVC. The decision was taken that the changes would be implemented as of the June 2018 administration of the CKE 2.

Appendix A

MODIFIED ANGOFF METHOD

WHAT IT IS → The Modified Angoff method of setting cut scores is the most popular method used with high-stakes examinations. With this method, experts evaluate each item on a test for difficulty and judge how likely it is that someone who is borderline in performance will get each item correct. Borderline candidates have, by definition, just enough competence to be considered competent (e.g., to pass the test). Any candidate showing the same or a higher level of performance as a borderline candidate is thus a "passing" candidate, and any candidate showing performance below the level of a borderline candidate is a "failing" candidate. The method has been successfully defended in court as being a fair method of setting cut scores that are used to make high-stakes decisions about candidates.

HOW IT'S DONE → The Modified Angoff method typically requires 5 to 15 experts in the field and is facilitated by a psychometrician. There are many variations of the Modified Angoff method used in practice, but generally the process begins with detailed training on how to apply ratings, followed by development of a description of the borderline candidate. Once training is complete (including a calibration exercise to make sure all raters have fully grasped the method), ratings are applied individually by each rater and compiled by the psychometrician. Discrepancies across raters are identified and flagged for discussion. Raters then have an opportunity to discuss their ratings and to rerate any items if the new information is considered cause to do so. In some cases, the psychometrician will introduce data from previous administrations of the item to further refine judgments. Once all items have been rated, an average Angoff rating for the exam is calculated by simply taking the average of all item ratings. The result is the cut score for the exam as a whole.

WHY IT'S USED → The benefit of the Modified Angoff method is that the resulting cut scores set an objective hurdle for candidates. Candidates who demonstrate performance above the borderline level (as systematically established by experts) are considered to have sufficient competence, and those below that level are considered to have insufficient competence. The proportion of candidates deemed below or above the cut score is not arbitrary and depends only on the actual ability of those candidates. For examinations resulting in pass/fail decisions, the implication of this is that all candidates would pass if they all showed better than the minimal accepted level of competence (i.e., above the borderline), or they would all fail if they all showed less than the minimal accepted level of competence. What is important is whether each candidate scores above or below the cut score, with that cut score being set based on the actual difficulty of the test and the expected performance of candidates showing the lowest level of acceptable performance. Because of this, the Modified Angoff method fairly assesses individual candidates on their own merits.

References

Cizek, G.J, & Bunch, M.B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications.

Plake, B.S., & Cizek, G.J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and yes/no standard setting methods. In G.J. Cizek (Ed.), *Setting performance standards* (pp. 181–199). New York, NY: Routledge.

Smith, I.L., & Springer, C.C. (2009). Standard setting. In Institute for Credentialing Excellence, *Certification: The ICE handbook* (pp. 235–264). Washington, DC: Institute for Credentialing Excellence.

Appendix B

Requiring candidates to pass all sub-tests on a certification exam (aka, non-compensatory scoring of certification exams)

Claude Balthazard, Vice-President, Regulatory Affairs and Registrar, Human Resources Professionals Association

John Wickett, Lead consultant and Principal, Wickett Measurement Systems Inc.

The challenge	Directive to require candidates to achieve thresholds in nine functional areas, in addition to an overall threshold, before a pass result will be granted. A candidate who passes overall, but who fails just one of nine functional areas, will fail and must retake the entire test.
The facts	1. Brand new high-stakes certification exam. 2. Exam with 225 scored four-option multiple-choice items. 3. Each functional area has 18 to 31 items, depending on blueprint weight.
The issues	1. Pass/fail decisions will need to be made based on subscores with as few as 18 items. 2. Decisions need to be defensible and candidate appeal must be anticipated.
What we did	1. Standard two-round Modified Angoff with eight judges conducted after initial administration. 2. Overall pass mark established using mean of all Angoffed values, with no adjustments. Pass mark was 138.5 out of 225, yielding a pass rate of 68.8%. 3. To calculate threshold for each functional area: a. Calculate the conditional standard error of measurement around the mean Angoff value for the functional area using the Lord method.[1] b. Multiply the CSEM by 2.417 to provide 95% one-tailed confidence across all nine comparisons.[2] This is equivalent to 99.22% confidence for each independent comparison. c. Subtract the resulting value from the mean Angoff value for the functional area. d. Use the rounded-up integer of this resulting value as the cut score for that functional area. 4. Based on only the functional area thresholds, nine additional candidates failed the exam. Thresholds ranged from 30% to 50% across functional areas, well below the mean performances (ranging from 57% to 73%).
What this accomplished	1. Candidates cannot pass the examination if they are ***substantially*** unknowledgeable in any one area. The format forces candidates to be generalists to at least some extent and not rely on strengths in a few areas. 2. Candidates who know their stuff across the board, with no areas of extreme weakness, will pass . . . exactly in line with the goals of the program.
Considerations for others	1. Consider explicitly how pass/fail decisions will be prioritized. a. In this case, for the overall score, a balance was struck where errors on either side of the pass mark were balanced. b. For the functional area thresholds, however, the priority was placed on ***not*** failing someone based on any one function area unless we were more than 95% sure. 2. The functional areas all had lower reliabilities (.44 to .71) than the overall score (.92), but this was accounted for by the CSEM adjustment. So while it is true that making decisions solely on subscores with so few items would be problematic, doing so in conjunction with an appropriate overall score pass mark may help achieve program goals.

[1] Feldt, L.S., Steffen, M. & Gupta, N.C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement, 9, 251-361.*

[2] Gupta, S.S. (1963). Probability integrals of multivariate normal and multivariate t. *The Annals of Mathematical Statistics, 34, 792-828.*

