PUBLIC RELEASE

Technical Report: June 2018 CKE 2

Human Resources Professionals Association

3 August 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 276 candidates using computer-based testing at Prometric test centres June 25–July 10, 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 220–230 best-performing items (while adhering to the prescribed distribution across functional areas). The mean score for first-time candidates (n=233[2]) was 152.2 (66.4%), and for all candidates it was 148.7 (64.9%), out of 229 scored items. 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 March 2018 and June 2017 CKE 2 administrations, yielding an integer pass mark of 133. 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 77.7% and a pass rate for all candidates of 72.7%.

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, who were identified as being statistical outliers, or who had written an alternative test 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 form (using a combination of scored and experimental test items). Wickett constructed the final test form according to the following parameters:

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

Wickett proofed the final form for text errors and detection of potential enemy items. Items flagged as enemies were replaced.

The final form composition for the primary June 2018 CKE 2 form is shown in Table 1. All functional areas are within 1 item of their targets, and as such, the form reflects the blueprint (see Appendix A for the CKE 2 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 candidate scores.

Table 1: Test form as administered

	Functional Area	Actual Items	Target	Variance
10	Strategy	27	27–28	—
20	Professional Practice	28	27–28	—
30	Organizational Effectiveness	34	35	−1
40	Workforce Planning & Talent Management	35	35	—
50	Labour & Employee Relations	23	22–23	—
60	Total Rewards	25	25	—
70	Learning & Development	26	27–28	−1
80	Health, Wellness & Safe Workplace	21	20	+1
90	HR Metrics, Reporting & Financial Management	31	30	+1
	TOTAL	**250**	**250**	**—**

Testing Window

The examination was administered via computer-based testing at Prometric test sites primarily in Ontario. The testing window was June 25–July 10, 2018, inclusive, and 276 candidates wrote the exam.

Candidates had access to a basic-function calculator on screen. No other aids or resources 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, each 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 problems were encountered.

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, 49 minutes (10 minutes more than in March 2018). The time limit on the CKE 2 was 5 hours, suggesting that time was not a factor in scores across candidates.

Sixteen candidates (6%) took the full 5 hours, suggesting that those candidates may have wanted more time, and 3 candidates (1%) left at least 1 item blank, suggesting that those candidates timed out of the exam before being able to complete it. These metrics will continue to be monitored, but at 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 no relation between time spent on items and performance. (Note that 4 candidates exceeded the time limit; these candidates 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 3 days and the last 4 days was an increase of 4.0 marks out of 250. This 1.6% change is not considered meaningful.

As a matter of interest, candidate volumes were also examined across the window; these are also shown in Figure 2. The usual pattern of increased sittings at the end of the window was generally in evidence, though tempered by the fact that the window stretched across the Canada Day holiday.

Figure 1: Examination time distribution for all candidates

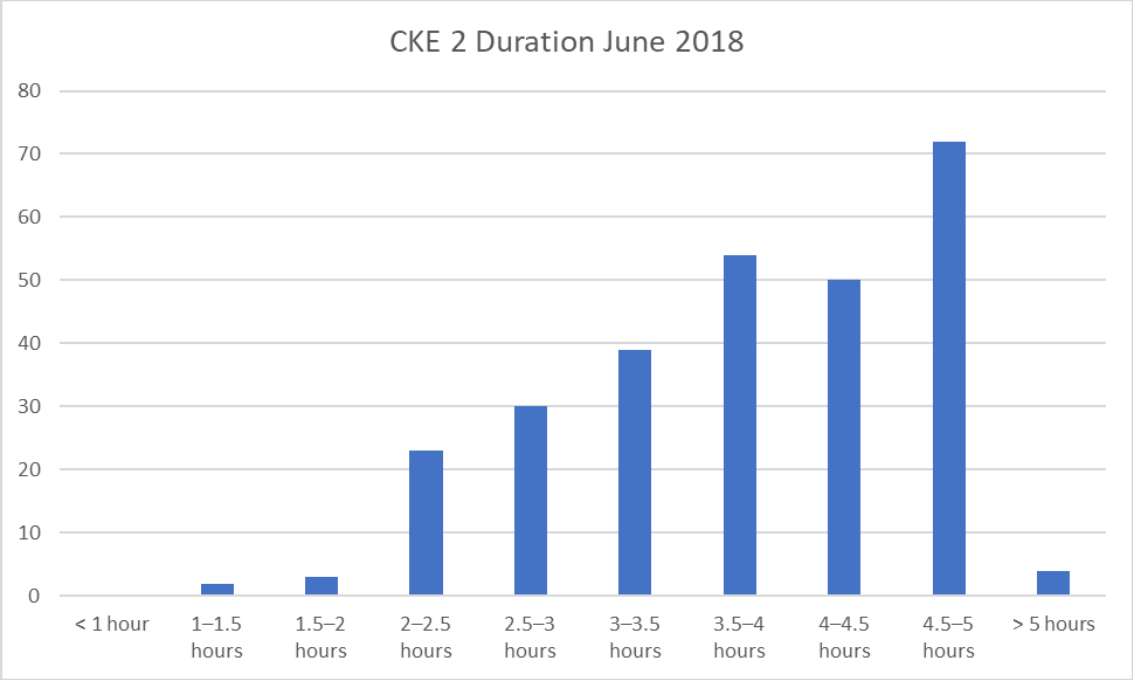
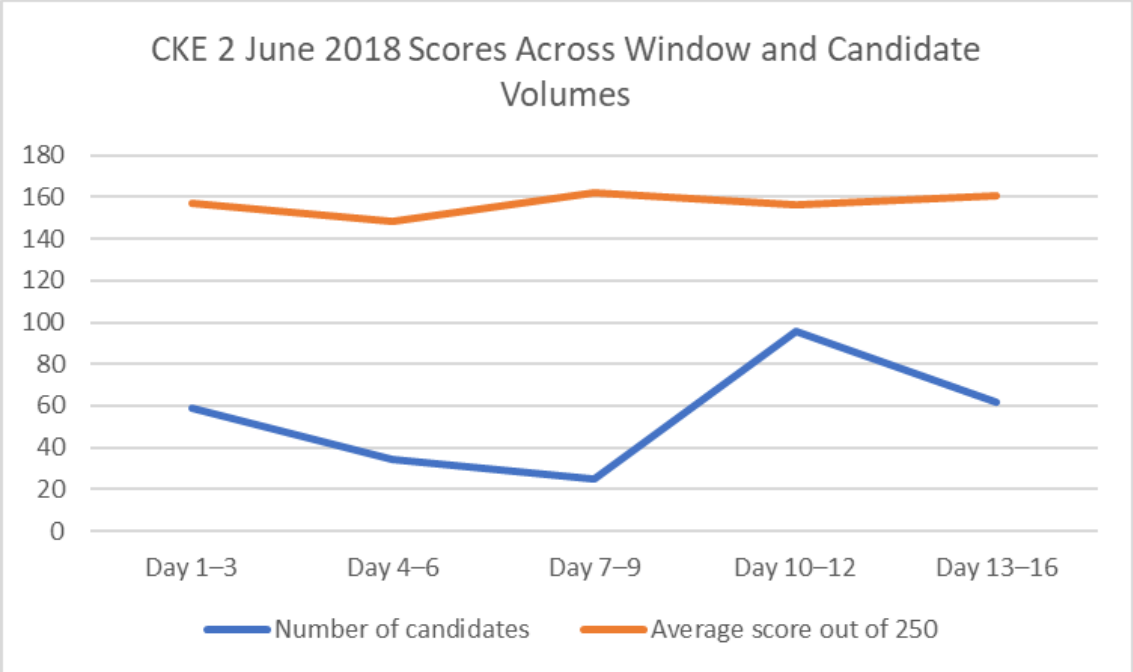


Figure 2: Candidate volume and score trends across testing window



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 calculated for all remaining candidates based on the full set of 250 items. One candidate was 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 subscores 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; candidates with abnormal response patterns were also removed. Candidates who left more than 5 blanks were also removed from analysis. 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, 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, 8 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 (54.1% and 64.0%, respectively, on the full exam of 250 items). This difference was meaningful and significant ($t(59)=7.28$, $p<.001$). In keeping with standard procedures, these candidates were removed from subsequent analyses. The CKE 2 analysis proceeded with 233 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; 270 responses were obtained from candidates (response rate, 98%).

Table 2 shows the responses to the administration-related questions. Note that candidates were generally very positive about the administration experience. Table 3 shows the content-related questions; there was a tendency to more neutrality on these questions. The 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 using a 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.	97	95	24	33	21	3.8	71%
2.	I was well informed about what documents to bring to the exam location.	155	102	7	4	1	4.5	96%
3.	Proctors enforced the exam-day rules and the security procedures at the test centre were what I expected.	161	88	11	6	2	4.5	93%
4.	Proctors were professional and courteous.	166	87	10	4	1	4.5	94%
5.	The tutorial helped me understand how to complete the examination on the computer.	143	114	9	2	0	4.5	96%
6.	Navigation through the examination was easy and intuitive.	155	98	10	3	2	4.5	94%

*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.	140	104	13	5	6	4.4	91%
8.	Information available prior to exam day provided me with adequate details about the content and format of the exam.	50	102	58	39	19	3.5	57%
9.	I feel I was adequately prepared to write this examination.	16	80	103	59	9	3.1	36%
10.	The questions in the examination were clearly written.	12	120	81	44	11	3.3	49%
11.	The terminology used in the examination was accurate.	13	132	94	23	5	3.5	54%
12.	The situations presented in the examination were realistic.	24	165	60	16	3	3.7	71%
13.	The questions in the examination reflected the examination blueprint.	13	85	114	39	10	3.2	38%
14.	The examination was a fair assessment of my ability.	9	71	100	66	20	2.9	30%

*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.	47	18%
I feel that completing the examination on a computer decreased my performance.	55	21%
I feel that completing the examination on a computer had no effect on my performance.	165	62%

Initial Analysis

The full CKE 2 examination was 250 items, of which approximately 225 were to be scored. The other 25 items were designated as experimental. 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 approximately 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
Total candidates	275
Candidates in analysis	233
Mean	160.4 (64.2%)
Range	89–211 (35.6–84.4%)
Cronbach's alpha	.93
Mean r_{pb}*	.22

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 June 25, 2018, using members of the CHRL Examination Validation Committee (EVC). The EVC (Table 6) was first reminded of basic item and test analysis methods and was oriented to the main statistics used to evaluate the quality of the CKE 2.

Note that there were fewer than the targeted 4 members on this call. Though this was not ideal the discussion on each item was still robust and did not reveal any lack of consideration.

Table 6: CHRL Examination Validation Committee members – Key validation

Member	Credential	Years of Relevant Experience	Years 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	Healthcare
Lisa Scian	CHRL	15–19	1	Information & communication technology

✓Participated in the session.

The group was informed that test reliability, as measured by Cronbach's alpha, was .930 based on the set of 250 potentially scored items and that this was above the generally accepted threshold of .80.

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 they were given less direction on those items with borderline statistics. Where available, candidates' comments about the items were also shown. The group made decisions based on content and the data through discussion; they removed 21 items that they felt were least appropriate to retain for scoring. Past item data were also used where available, and the group was directed to consider these data as an

addition to statistics from the June administration. Panel members' comments 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. Most were moderate to strong items, however. The final alpha for the set of 229 scored items was .934. The difficulties ranged from 25.3% to 94.8%, with a mean of 66.4%. The r_{pb}* values ranged from .00 to .49 with a mean of .24.

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

The group endorsed the final set of items for use in scoring the June 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**	24	21	25	29	11% ± 2%
20	**Professional Practice**	24	21	25	29	11% ± 2%
30	**Organizational Effectiveness**	30	28	32	36	14% ± 2%
40	**Workforce Planning & Talent Management**	31	28	32	36	14% ± 2%
50	**Labour & Employee Relations**	21	17	21	25	9% ± 2%
60	**Total Rewards**	24	19	23	27	10% ± 2%
70	**Learning & Development**	26	21	25	29	11% ± 2%
80	**Health, Wellness & Safe Workplace**	21	14	18	22	8% ± 2%
90	**HR Metrics, Reporting & Financial Management**	28	23	27	32	12% ± 2%
Total		**229**				

Establishing the Pass Mark: Equating

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

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

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 B. Specific information on the standard-setting session is provided in the Technical Report issued for the November 2015 administration.

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 9 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 March 2018). The intention following these 2 equating runs was to average them to arrive at a final pass mark for the June 2018 CKE 2. These administrations were chosen because they were the most recent administration and the administration corresponding roughly to the same administration month the previous year.

Equating Back to the March 2018 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 3 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. Items with an increase or decrease of 10% in terms of difficulty were also removed as anchors. 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 .24 (for June 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 March 2018

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

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

The mean, Tucker, Levine observed-score, and circle arc 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 June 2018 CKE 2 scored somewhat better than the population taking the March 2018 CKE 2 (66.8% vs. 65.7%, respectively; $t(374)=0.88$, *ns*). Because the June 2018 CKE 2 candidates scored somewhat better (based on the anchors), they would likely have a higher pass rate than was seen in March.

The equating analysis bears this out (Table 10) for the most part. The Tucker, Levine observed, circle arc, and mean methods indicate a pass mark of 133–134. The pass rate based on this equating run is somewhat higher than what was seen in March 2018. The Tucker equating value of 133.17 was extracted from this analysis for use in setting the final pass mark.

Table 9: Equating parameter table – Total pass mark, to March 2018

		March 2018	June 2018
	n	143	233
	Scored items	225	229
Mean score	Total	67.0%	66.4%
Mean score	Anchors	65.7%	66.8%

Table 10: Equating outcome table – Total pass mark, to March 2018

	Pass Mark		Pass Rate	
Method	**Precise**	**Integer**	**All**	**First-time**
Combo March 2018	**134.77**	**135**	**64.1%**	**72.7%**
Tucker	133.17	134	*69.1%*	*75.5%*
Levine observed	132.99	133	*72.7%*	*77.7%*
Circle arc 1	133.54	134	*69.1%*	*75.5%*
Circle arc 2	133.51	134	*69.1%*	*75.5%*
Mean	133.96	134	*69.1%*	*75.5%*

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 3 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. Items with an increase or decrease of 10% in terms of difficulty were also removed as anchors. 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 0.66 and a mean corrected point-biserial of .25 (for June 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%	11%
20	10%	11%
30	14%	14%
40	13%	14%
50	10%	9%
60	10%	10%
70	11%	11%
80	10%	8%
90	12%	12%

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

The mean, Tucker, Levine observed-score, and circle arc 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 June 2018 CKE 2 scored somewhat higher than the population taking the June 2017 CKE 2 (66.5% vs. 65.6%, respectively; $t(430)=0.75$, *ns*). Because the June 2018 CKE 2 candidates scored somewhat higher (based on the anchors), they would likely have a somewhat higher pass rate than was seen in June 2017.

The equating analysis bears this out (Table 13). The Tucker, Levine observed, circle arc, and mean methods indicate a pass mark of 133–135. The pass rate based on this equating run is somewhat higher than in June 2017, as expected. The Tucker equating value of 132.53 was extracted from this analysis for use in setting the final pass mark.

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

		June 2017	June 2018
	n	199	233
	Scored items	225	229
Mean score	Total	65.5%	66.4%
Mean score	Anchors	65.6%	66.5%

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**	**70.0%**	**75.9%**
Tucker	132.53	133	*72.7%*	*77.7%*
Levine observed	132.36	133	*72.7%*	*77.7%*
Circle arc 1	133.81	134	*69.1%*	*75.5%*
Circle arc 2	133.81	134	*69.1%*	*75.5%*
Mean	134.24	135	*68.4%*	*75.1%*

Combined Results

Table 14 shows the pass mark values across the 2 equating runs. The value highlighted in green 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 132 and 134. Barring a clear reason to choose otherwise, the simple arithmetical mean (132.848799) of the 2 identified values was the recommended pass mark for the June 2018 CKE 2.

Using the established convention for this testing program, the mean combined value was rounded up to a cut score of 133. The resulting pass rate of 77.7% for first-time candidates is somewhat higher than seen in June 2017 and March 2018 (see Table 15), but there was evidence that the June 2018 candidates had greater ability and so this rate was consistent with expectations. The pass rate for all candidates was 72.7%.

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

	Mar. 18	June 17
Tucker	133.2	132.5
Levine observed	133.0	132.4
Circle arc 1	133.5	133.8
Circle arc 2	133.5	133.8
Mean	134.0	134.2

Table 15: Historical pass rates – Total pass mark

	All	First-time
June 16	73.8%	78.5%
Nov. 16	64.6%	68.1%
Mar. 17	59.4%	72.9%
June 17	70.0%	75.9%
Oct. 17	69.3%	75.7%
Mar. 18	64.1%	72.7%
June 18	**72.7%**	**77.7%**

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 C for a conference presentation 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 June 2017 and March 2018 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 sufficiently achieved.

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 1 candidate failed the exam based solely on having missed the threshold for a functional area.

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

Table 16: Alignment between difficulty of anchors and full exam

Area*	March 2018 Anchors	June 2017 Anchors	Full Exam
10	69%	71%	69%
20	67%	61%	63%
30	67%	66%	66%
40	65%	67%	66%
50	67%	69%	67%
60	68%	65%	67%
70	67%	68%	68%
80	67%	67%	67%
90	64%	64%	65%

*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 Mar. 18	Tucker	10.17	6.98	10.64	9.74	7.09	9.08	10.13	6.62	9.28
To June 17	Tucker	9.77	7.05	10.54	8.96	7.82	8.28	10.84	6.69	8.71
	Average	9.97	7.01	10.59	9.35	7.46	8.68	10.48	6.66	9.00
	Integer	10	8	11	10	8	9	11	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	**9.97**	10	24	42%	36%	.61	16.5	**0**
20	**7.01**	8	24	29%	42%	.55	15.1	**0**
30	**10.59**	11	30	35%	36%	.61	19.8	**1**
40	**9.35**	10	31	30%	34%	.73	20.5	**0**
50	**7.46**	8	21	36%	39%	.55	14.0	**0**
60	**8.68**	9	24	36%	39%	.58	16.2	**0**
70	**10.48**	11	26	40%	37%	.61	17.7	**0**
80	**6.66**	7	21	32%	39%	.53	14.0	**0**
90	**9.00**	9	28	32%	29%	.67	18.3	**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 would have passed at the total score level.

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

Fails	Both measures	28	10.2%
	Total score only	47	17.1%
	Functional area score only	1	0.4%
Passes	Neither	199	72.4%

Pass Mark Approval

The total score pass mark, the thresholds for all functional areas, and the process used to derive them were presented to the CHRL EVC (Table 20) via teleconference on July 26, 2018. The committee approved the process and cut scores (which were presented along with the consequent pass rate) for recommendation to HRPA. The HRPA Registrar accepted the recommendation from the committee on the same call, and the total and functional area cut scores were formally established.

Table 20: CHRL Examination Validation Committee members – Pass mark approval

Member	Credential	Years of Relevant Experience	Years 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	Healthcare
Lisa Scian	CHRL	15–19	1	Information & communication technology

✓Participated in the session.

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 9 functional areas were also computed for each candidate. An Excel file with the final candidate results was provided to HRPA.

Table 21 provides the means and standard deviations for the functional areas and for the total score, using all candidates who took the new June 2018 CKE 2 form. Table 22 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 21: Total and functional area scores for all candidates

	Functional Area	Percentage	Mean	SD*
10	**Strategy**	68%	16.2	3.3
20	**Professional Practice**	61%	14.6	3.3
30	**Organizational Effectiveness**	64%	19.2	4.0
40	**Workforce Planning & Talent Management**	65%	20.1	4.7
50	**Labour & Employee Relations**	65%	13.6	3.1
60	**Total Rewards**	66%	15.9	3.3
70	**Learning & Development**	66%	17.3	3.7
80	**Health, Wellness & Safe Workplace**	66%	13.8	3.0
90	**HR Metrics, Reporting & Financial Management**	64%	18.0	4.1
Total score		64.9%	148.7	26.3

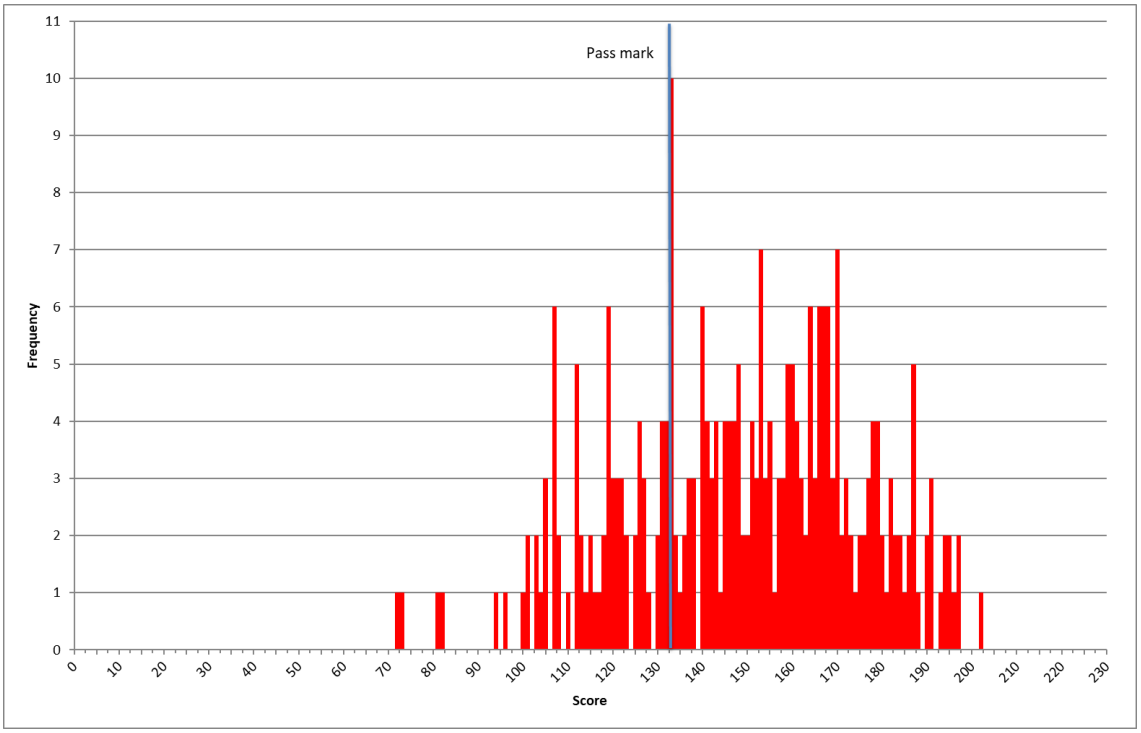
*SD = Standard deviation.

Table 22: Correlations between functional area scores for all candidates

Area*	10	20	30	40	50	60	70	80	90
10		.62	.65	.64	.58	.62	.67	.59	.62
20			.61	.64	.58	.60	.65	.57	.62
30				.69	.59	.62	.67	.57	.64
40					.59	.63	.73	.58	.58
50						.57	.63	.57	.55
60							.61	.55	.59
70								.61	.61
80									.50
90									

*See Table 21 for the full name of each functional area.

Figure 3: Score distribution for all candidates



Key Examination Metrics

Table 23 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 as of June 2018 the number of scored items was free to vary from 220 to 230 based on the number of experimental items and the work of the CHRL EVC during key validation.

Table 23: Key examination metrics – Candidates included in analysis only

Index	June 2018	March 2018	October 2017	June 2017	March 2017
Scored items	229	225	225	225	225
Candidates	233	143	235	199	70
Mean	152.2 (66.4%)	150.7 (67.0%)	148.5 (66.0%)	147.3 (65.5%)	151.3 (67.2%)
Median	155 (67.7%)	152 (67.6%)	150 (66.7%)	150 (66.7%)	156 (69.3%)
Skewness	−0.399	−0.132	−0.403	−0.295	−0.498
Kurtosis[i]	−0.424	−0.555	0.108	−0.306	−0.692
Range	81–202 (35.4–88.2%)	88–204 (39.1–90.7%)	78–203 (34.7–90.2%)	78–195 (34.7–86.7%)	93–188 (41.3–83.6%)
Standard deviation	25.22	24.53	22.98	23.67	23.91
Cronbach's alpha	.93	.93	.92	.93	.93
Mean r_{pb}*	.24	.23	.21	.23	.23
SEM[ii]	6.48	6.43	6.52	6.34	6.34
SEM at the pass mark	6.96	6.89	6.94	6.74	6.79
Decision consistency (uncorrected)[iii]	.93	.89	.90	.90	.92
Perceived fairness[iv]	30%	27%	28%	27%	31%
Pass mark	132.849	134.772	132.371	131.061	134.664
Effective pass mark	133	135	133	132	135
Pass rate	77.7%	72.7%	75.3%	75.4%	72.9%

[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

No development activities were conducted for the CKE 2 since the last administration of the CKE 2 in March 2018.

Appendix A

Blueprint

Comprehensive Knowledge Examination 2

Human Resources Professionals Association
Version 2.0

Approved by CHRL Exam Validation Committee March 13, 2018
Approved by HRPA Registrar March 14, 2018
Effective June 2018 administration

Credentials

Passing the Comprehensive Knowledge Examination 2 is a requirement for certification for CHRL candidates. The examination reflects the *HRPA Professional HR Competency Framework* (2014).

Purpose

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.

Structure

The structural variables provide high-level guidance as to what the examination will be like.

Table 24: CKE 2 Blueprint structural variables

Item types	Independent 4-option multiple choice
Length	250 items in total
	20–30 experimental items
Duration	Up to 5 hours
Delivery mode	Computer-based testing in proctored test centres
Frequency	3 windows per year

Content Weighting

The functional area weights were set in 2014 through a national survey and modified slightly in 2018 to remove weighting for competencies most appropriately tested on the CHRL Employment Law Examination. Within each functional area, items are distributed roughly evenly across the related competencies.

Table 25: Functional area weights on the CKE 2

Functional Area		CKE 2	
		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%

Table 26: Competencies not eligible on the CKE 2

FA	Comp
20	C035
	C036
	C037
50	C117
60	C139
80	C177
	C179
90	C204
	C205

Appendix B

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.

WICKETT

Appendix C

Human Resources Professionals Association

HRPA

ICKETT

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.*

