Summary Report on the Hypothetical Certification Program's 2016 Specialty Examination Administration

February 2017

Prepared by:
Julian Consulting

This version is NOT Confidential

Summary Report Table of Contents

Summary Report on the 2016 Administration of the Hypothetical Certification Program's Specialty Examination

Introduction

This report summarizes the development, administration, scoring, and results of the 2016 Specialty examination of a Hypothetical Certification Program (HCP). Overall, the form performed well, with passing rates similar to those of previous years.

Test Design Description

HCP examinations are administered in English once a year in Vendor's computer-based testing centers, during an approximately two-week-long testing window. A single form is built each year with approximately half new items and half previously-used items selected from the item bank. The used items' bank difficulties form an equating link to previous exams, allowing the passing standard set in 2013 to be applied to the new form. Item responses are analyzed with the Rasch Model after test administration and the final set of high-performing, content-representative items are selected for scoring. Using the bank difficulties of the used items, the cut score (number-correct score on this test form that is equivalent to the passing standard on the bank scale) for the final item set is determined and applied to examinee scores. Test results are emailed to examinees approximately 60 days after the end of the testing window.

Content Outline

The Job/Task Analysis (JTA) on which the current content outline is based was conducted in 2012. Policies require a JTA every five years, so the next one will be conducted in 2017. For details on the process used in 2012 and the demographics of the respondents and decision makers, see the JTA final report. The content outline is included in Appendix A.

Passing Standard

After the implementation of the content outline resulting from the 2012 JTA, a standard-setting study was conducted in 2013 (see Standard Setting final report for details). The 2013 passing standard was equated to the 2016 forms through common-item Rasch equating.

Test Development

Item Writing and Review Results

The Specialty examination committee enhanced their item banks by approving 163 new items to the content areas of greatest need – mostly Domain 7. An additional 20 new items were created through committee editing of previously-used items that did not perform well. All items were multiple-choice, single-best-answer questions.

Two-hundred and three (203) new items were written remotely by the committee members using the item banking system and 50 using the old Microsoft Word templates, which will be eliminated next year because of the extra work they create for staff.

The 2015 examination's item analysis identified 40 previously-used items for review and revision based on examinees' responses. The total of 243 items was reviewed through two different mechanisms, one in-person and one virtual.

A total of 133 new and used items were reviewed at an item-development workshop held on the day before the national Hypothetical Convention, in February 2016. The 12 reviewers represented a cross-section of hypothetical practitioners in geography, size of practice, and demographics. They approved a total of 133 items - 50 items as written, and 83 with edits. They set aside 30 for later work, and rejected 20, with "triviality" being the most commonly-cited reason.

For the online item-review workshop on April 15, 2016, a dozen invitations were issued to previous participants at an in-person item development workshop. Eight participated, and again their demographics represented the population. A video-conference was held and, after reminding volunteers about security protocols, the Exam Program Manager led the group through the review of the remainder of the year's new and rework items. During the two-hour review session, they approved ten items as written and 20 with edits, set aside 10 for later work, and rejected 5.

Item Bank Sufficiency for Form Building

The item bank's status at the beginning of form assembly is shown in Appendix A. This includes the 163 newly reviewed and approved items. Several domains are chronically under-represented in the bank, especially the largest domain, Domain 7. Note that some items are not assigned to a content category.

The Content Outline controls only domain percentages, but encouraging the writing of items to sparsely populated subdomains can result in better coverage and easier identification of item enemies than allowing volunteers to choose their own topics.

Selection of Items for Forms

Items on Specialty exams come in two types: new and used. Used items have statistics from prior use. New items do not. By policy, Specialty forms can have between 40% and 60% new items. The used items constitute the equating link between this test form and the item bank's difficulty scale. It will determine the equating adjustment needed to put the new items onto the bank scale, and it allows application of the passing standard, which is on the bank scale, to be applied to the new form.

Preliminary test forms matching the current Content Outline (see the rightmost column in Appendix A) were assembled by Vendor with an approximately 10% overage of items (at least one extra item) in each content area, for a total of 115 starting items. HCP exam committee members met via conference call to select the final set of 100 content-balanced items for inclusion on the 2016 test forms.

Vendor staff created the XML version of the items, appended the welcome screen, NDA agreement, tutorial, and after-exam survey, and published it to the Vendor's test-administration software platform on May 10, 2016.

Test Administration

Application for the HCP credential was handled by the HCP office via an online process. Supporting documents were supplied by the applicant and forwarded to HCP judges to determine eligibility for the credential. The list of those individuals deemed eligible was forwarded to Vendor. At the end of the registration period, this list included 75 individuals. See Appendix B for registration counts by state. The candidates selected a testing site, date, and time via the Vendor online system or by telephone.

A total of 105 Specialty examinations were administered in 350 Vendor computer-based test centers from November 27 to December 15, 2016, with an extension to December 29, 2016 because of scheduling issues. Counts of examinations administered on each day are included in Appendix C.

No incidents were reported during test administration beyond simple restarts of a computer workstation. All candidate identification was adequate, and all score reports were delivered.

Candidate responses to Vendor's survey questions about their testing experience are shown in Appendix D. Note the high percentage of candidates dissatisfied with {something}. This may warrant further investigation with Vendor. The candidates' comments are in Appendix E.

Candidate Demographics

The candidates self-identified their specialty and training information in the after-exam survey. The results are shown below, with the largest group indicated by shading:

Table 1. Candidate Demographics

		Oneology	Twoology	Threeology	Fourology	Total
Training	**Count**	87	3	2	13	105
	% of Respondents	**82%**	**3%**	**2%**	**12%**	
		East	**Midwest**	**South**	**West**	**Total**
Region of Country	**Count**	23	42	25	15	105
	% of Respondents	**22%**	**39%**	**23%**	**15%**	
…more questions						

ADA Accommodations

For the 2016 examination, HCP received 13 requests for ADA accommodations. Ten of the requests were granted (8 for extra time and 2 for a separate room), and three were denied for lack of documentation.

Scoring and Results

Selection of the Final Set of Scored items

Item analyses after test administration revealed that some items were too difficult or too easy to contribute to the measurement of candidates, and other items showed low discrimination-values (point-biserial correlations with total test scores). Such items are typically removed from scoring. Items with low point-biserials are often ones measuring skills unrelated to the underlying construct of the test. Such items actively detract from the measurement precision of the examinations. HCP policy requires the removal of items with negative point-biserials after review by the Exam Development Committee. Any extremely difficult items (p-value<.30) or items with low (<.20) point-biserials will be brought to the Exam Development Committee for review and a decision about inclusion in scoring.

Four items with negative point-biserials or extreme difficulties were proposed to the Exam Development Committee for removal from scoring during their July 13, 2016 conference call. Three were sent to rework for next year. One was deleted entirely. Five difficult items and three with low positive point-biserials were reviewed and deemed correctly-keyed by the Committee, and allowed to remain in the scored item set.

Table 2. Item Performance on 2016 Specialty Examination by Content Area

Specialty 2016 Exam	Administered Items	# items scored	Scored %	Content Outline %	Difference**	Average p-value*	Average Pt. Bis*	Median Seconds *
Domain 1	7	7	7%	7%	<2	0.64	0.28	55
… other domains								
Grand Total	**100**	**96**	**100%**	**100%**		**0.68**	**0.18**	**59**

*Of the final scored items
**HCP policy is that scored item proportions will differ from the Content Outline by no more than two percentage points.

Test Reliability and Decision Consistency

Two types of reliability statistics are reported. Alpha is a commonly-cited reliability statistic for norm-referenced tests, assessing whether people would rank order in the same way across multiple sets of items or administrations of the exam. On the scale from zero to 1.0, a value of .90 or higher is generally considered desirable for high-stakes examinations; a value of .80 is acceptable. Specialty's alpha reliability was .81. Test length and candidate variability both influence this statistic. The average point-biserial coefficient directly impacts the reliability, as well; so removing the items with low or negative point-biserials also improves reliability. This .81 is on the low side of acceptable, suggesting a need for more discriminating items.

Livingston's Coefficient is concerned with the consistency of pass/fail decisions and, therefore, may be a more appropriate choice for evaluating the quality of the HCP examination. Specialty had a Livingston's Coefficient of .87. This is a high value, indicating a satisfactory level of decision-making precision for a high-stakes examination. The high decision-consistency index suggests that the examination form's consistency is targeted in the vicinity of the passing standard, where it's most needed. This implies that the overall difficulty of the items is good.

Calibration of Items and Candidates with the Rasch Model

Percent-correct scores are impacted by both the ability of the candidates and the difficulty of the items, so these scores cannot be directly compared across tests or years. Item Response Theory (IRT) scores enable longitudinal comparisons and item banking, with all items having difficulties on the same scale, regardless of the ability of its candidates.

Specialty's items and candidate were calibrated with the Rasch Model using Winsteps calibration software to estimate item difficulties and candidate abilities and equate them to the item bank scale. The average Specialty candidate answered 67% of the scored items correctly, which is equivalent of an average ability on the bank scale of 1.18.

Figure 1 shows the distribution of percent-correct scores on the Specialty examination. This year's distribution for Specialty appears to be about the same symmetrically as in the 2015 distribution. Table 3 reports summary statistics for the candidates, the items, and the test as a whole.

Figure 1. Distribution of Percent-Correct Scores

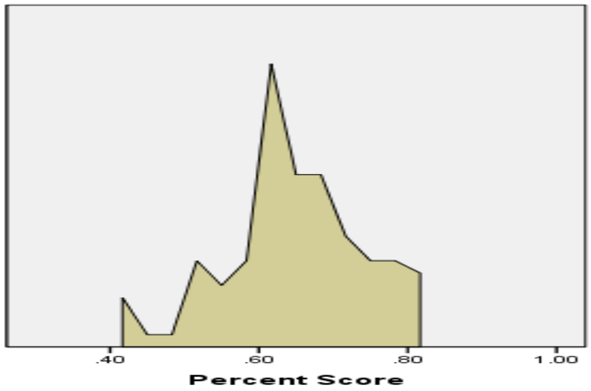


Table 3. Summary Data for 2016 Specialty Examination

2016 Specialty Exam	Statistic	N	Min	Max	Mean	SD
Candidate Information	Number Correct Score	105	41	82	64.70	9.66
	Proportion Correct Score	105	0.43	0.85	0.67	0.10
	IRT Score	105	-0.11	2.36	1.18	0.56
	Test Duration (in mins)	105	72.27	120.62	112.74	12.59
Item Information	Point-Biserial of active items	96	-0.07	0.42	0.18	0.11
	P-Value of active items	96	0.13	0.95	0.68	0.18
Test Information	Alpha Reliability Coefficient				0.81	
	Livingston's Coefficient				0.87	

Equating the 2013 Passing Standard to this Form

In October 2013, the HCP Test Development Committee determined the passing standard for the Specialty examination. The passing standard of 0.74 on the item bank scale was used for the 2014, 2015, and 2016 examinations. The passing standard of 0.74 on the bank scale translated to 57 of the 96 scored items (59%) correct. It has ranged from 55% to 65% over the past several years, as test forms vary in difficulty (see Table 4).

Table 4. Exam Performance Across Time

	2012	2013	2014	2015	2016
Number of Candidates	108	102	136	119	105
Avg. Percent Correct	75%	72%	68%	69%	67%
Number of Items Scored	95	95	94	97	96
Avg. IRT Score	1.38	1.22	1.09	1.21	1.18
Bank Scale Passing Standard	0.66	0.66	0.74	0.74	0.74
Cut Score Percent Correct	54	55	59	60	59
Avg. Time Duration (in mins)	110	114	112	112	113

Point-Biserial of Active Items	0.20	0.19	0.20	0.17	0.18
Cronbach's Alpha Reliability	0.79	0.84	.80	.80	.81
Livingston Coefficient	0.83	0.87	0.86	0.86	0.87

Passing Rates by Demographic Group

Score reports were mailed 60 days after then end of the testing window. Of the 105 Specialty candidates, 86 passed (an 81% pass rate). Performance of all demographic groups was essentially identical (see Table 5).

Table 5. Passing Rates by Candidate Demographics

		Oneology	**Twoology**	**Threeology**	**Fourology**	**Total**
Training	**Count**	87	3	2	13	105
	% PASSing	**67%**	**66%**	**68%**	**67%**	
		East	**Midwest**	**South**	**West**	**Total**
Region of Country	**Count**	23	42	25	15	105
	% PASSing	**67%**	**66%**	**68%**	**67%**	

Score Scale for Reported Scores

Before sending scores to candidates, the bank-scale scores are converted to a score scale with which candidates are familiar and that retains a constant meaning across time. For example, a 500 is always the lowest passing score on the scale, whereas it might be 55% one year and 58% in another. The bank scale is used as the candidate's original score, not the percent correct, because the impact of item difficulty has been removed from it by the Rasch Model.

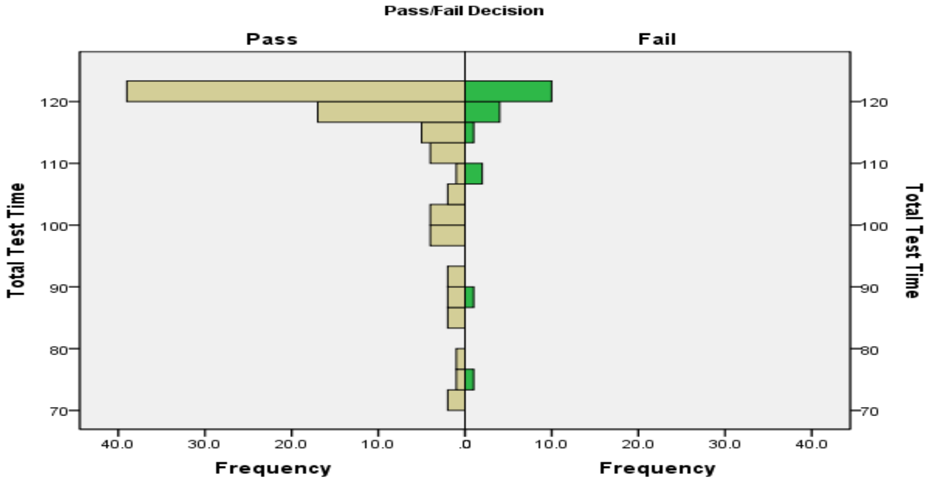
Test Speededness

On average, candidates took close to the maximum time, 113 of 120 allowed minutes (Table 3). Figure 2 shows the distribution of minutes spent on the examination. Note the large number of people who spent the maximum allowed time. A general rule of thumb is that 5% of candidates will stay until the end, no matter how generous the time. On Specialty, 44% used it all.

The HCP needs to determine whether they intend to include speed reading and answering questions is an intentional part of their construct. If not, speededness will need to be reduced. Average time spent on an item can be used in item selection to control test speededness.

However, using all the time is not associated with lower scores. The correlation between time spent on the exam and score was not statistically significant. Figure 2 shows that the spread of people across time is similar for both Pass and Fail groups.

Figure 2. Distribution of Time Spent on Exam



Recommendations

1. Content Outline Review in 2017, Item Bank Review, and Item Writing Needed

After five years with the same basic content outline, it is time for a Job Task Analysis survey to determine if the current content outline is still fresh and relevant to current practice. Also, the item counts shown in Appendix A includes ***all*** items, even those determined to be poor performers. **An item bank review** of all items would be beneficial to weed out older, less viable items.

As recommended in 2015, our eventual **goal** for the HCP examination programs is to have **at least two forms of each examination, overlapping by approximately 25%** of the items, and for each year's forms to overlap with the previous year's by 25%. But, while the candidate volumes are still low, a single form each year will suffice. Each year's single form, however, should be substantially different (with up to 75% new items) from the previous year. The **2016 form had only 40% new items**. Therefore, we recommend developing at least a form's worth of new items in each content area.

2. Item Development Workshop Again in 2017

The in-person Item Development Workshops have been a success. The number of items approved and edited **enhanced the item banks so that item selection was easier** than ever before. For 2017, we **recommend another in-person workshop** for editing and reviewing exam items. Volunteers are also needed to p**rovide items' missing content classifications**.

3. Research Needed on Speededness, Remote Item Writing, and the new Content Outline

HCP needs to decide if the **ability to work quickly** is part of what they're trying to measure. If not, revise item development and selection processes to remove items that take too long from the pool.

The several available ways to produce items (in-person and remote, new and revising), will be investigated for cost versus productivity.

2017 will be the year of the JTA, with a new content outline to be used for the 2018 test form. Substantial content changes may require an additional year of item development, so don't advertise the go-live date until the JTA results are determined.

4. Keep Passing Standard for 2017

It is recommended that the passing standards set by HCP in 2013 and applied in 2014, 2015, and

2016 be **carried over to the forms again in 2017**, anticipating an **updated content outline in 2018, and, therefore, potentially a new passing standard**.

5. Standard-setting Panel May be Needed in 2018

If the content outline changes substantially, a **standard-setting panel** of a dozen SMEs may need to be convened right **after the 2018 test administration**.

Appendix A. Specialty Examination Content Outline and Item Pool Size

Specialty Content Outline	**# Items in Pool**	**% Items in Pool**	**% of Exam**
Domain 1	**25**	**6%**	**7%**
Needs subdomain classification	1		
Subdomain A	4		
Subdomain B	7		
Subdomain C	13		
Domain 2	**32**	**7%**	**4%**
Subdomain A	11		
Subdomain B	12		
Subdomain C	9		
Domain 3	**35**	**8%**	**8%**
Subdomain A	12		
Subdomain B	17		
Subdomain C	6		
Domain 4	**119**	**27%**	**30%**
Needs subdomain classification	3		
Subdomain A	27		
Subdomain B	43		
Subdomain C	46		
Domain 5	**27**	**6%**	**7%**
Subdomain A	8		
Subdomain B	19		
Domain 6	**44**	**10%**	**12%**
Subdomain A	17		
Subdomain B	9		
Subdomain C	7		
Subdomain D	11		
Domain 7	**116**	**26%**	**30%**
Subdomain A	39		
Subdomain B	60		
Subdomain C	17		
Domain 8	**20**	**5%**	**2%**
Needs Subdomain classification	3		
Subdomain A	8		
Subdomain B	9		
Needs classification	**23**	**5%**	
Total **NB: Counts include items that did not perform well on the test and items that need further editing.**	**441**	**100%**	**100%**

Appendix B. Candidate Registrations by State

State	Registrations
AL	1
AZ	2
CA	1
CO	2
FL	5
GA	2
HI	1
IA	2
IL	2
IN	3
KS	1
KY	1
LA	2
MA	1
ME	1
MI	9
MN	1
MO	2
MS	1
NJ	5
NM	1
NV	1
NY	4
OH	6
OK	1
PA	1
SC	1
SD	
TN	3
TX	4
VA	3
WA	1
WI	1
WV	1
Canada	11
Germany	3
Puerto Rico	1
United Kingdom	17
Total	**105**

Appendix C. Exams Administered by Test Date

	Count
Tuesday, 27-Nov-2016	1
Wednesday, 28-Nov-2016	1
Thursday, 29-Nov-2016	1
Friday, 30-Nov-2016	2
Saturday, 1-Dec-2016	2
Monday, 3-Dec-2016	3
Tuesday, 4-Dec-2016	3
Wednesday, 5-Dec-2016	6
Thursday, 6-Dec-2016	10
Friday, 7-Dec-2016	3
Saturday, 8-Dec-2016	6
Monday, 10-Dec-2016	5
Tuesday, 11-Dec-2016	3
Wednesday, 12-Dec-2016	5
Thursday, 13-Dec-2016	22
Friday, 14-Dec-2016	12
Saturday, 15-Dec-2016	17
Friday, 21-Dec-2016	1
Friday, 28-Dec-2016	1
Saturday, 29-Dec-2016	1
Total	**105**

Appendix D. Comparison of Statistics Across Years

Specialty	**2011**	**2012**	**2013**	**2015**	**2016**
Number of Candidates	108	102	136	119	105
Candidate % Correct	73%	69%	68%	72%	67%
IRT Score	1.38	1.22	1.09	1.21	1.18
Time Duration (in mins)	110	114	112	112	113
Point-Biserial of Active Items	0.20	0.19	0.20	0.17	0.18
Cronbach's Alpha Reliability	0.79	0.84	.80	.80	.81
Livingston Coefficient	0.83	0.87	0.86	0.86	0.87

Appendix E. Responses to Post-test Survey Questions

		Very Satisfied	**Satisfied**	**No Opinion**	**Dissatisfied**	**Very Dissatisfied**	**No Response**	**Total**	**Satisfaction Index***
Reservation process	#	56	63	11	24	20	1	175	**68%**
	%	**32%**	**36%**	**6%**	**14%**	**11%**	**1%**		
Test date and location	#	42	58	11	27	36	1	175	**57%**
	%	**24%**	**33%**	**6%**	**15%**	**21%**	**1%**		
Test center location	#	70	85	11	4	4	1	175	**89%**
	%	**40%**	**49%**	**6%**	**2%**	**2%**	**1%**		
	%	**44%**	**48%**	**5%**	**1%**	**1%**	**1%**		
…more questions	#								
	%								
Overall satisfaction with the testing process	#	51	89	18	9	7	1	175	**80%**
	%	**29%**	**51%**	**10%**	**5%**	**4%**	**1%**		

* Satisfaction Index is the number satisfied or very satisfied divided by the total number of respondents.

Appendix F. Candidates' Write-in Comments

Included as needed