ORIGINAL RESEARCH

Validating the Test Plan Specifications for the American Board of Family Medicine's Certification Examination

Thomas R. O'Neill, PhD, Michael R. Peabody, PhD, Keith L. Stelter, MD, MMM, James C. Puffer, MD, and John E. Brady, MD

Purpose: **To demonstrate the degree to which the American Board of Family Medicine's certification examination is representative of family physician practice with regard to frequency of diagnoses encounter and the criticality of the diagnoses.**

Methods: **Data from 2012 National Ambulatory Medical Care Survey was used to assess the frequency of diagnoses encountered by family physicians nationally. These diagnoses were also rated by a panel of content experts for how critical it was to diagnose and treat the condition correctly and then assign the condition to 1 of the 16 content categories used on the American Board of Family Medicine examination. These ratings of frequency and criticality were used to create 7 different new schemas to compute percentages for the content categories.**

Results: **The content category percentages for the 7 different schemas correlated with the 2006 to 2016 test plan percentages from 0.50 to 0.90 with the frequency conditions being more highly correlated and the criticality conditions being less correlated.**

Conclusions: **This study supports the continued use of the current Family Medicine Certification Examination content specifications as being representative of current family medicine practice; however, small adjustments might be warranted to permit better representation of the criticality of the topics.**
(J Am Board Fam Med 2019;32:876–882.)

Keywords: **Certification, Educational Measurement, Family Physicians, General Practitioners, Health Care Surveys, Medical Education, Psychometrics, Validity, Exam Blueprint**

The purpose of certification and licensure is to provide some degree of assurance to the public that an individual has met specific standards related to a scope of practice for a profession. Standardized examinations are a common method for permitting those who are entering a profession to demonstrate that they are competent to provide those professional services. Therefore, standardized examinations must adequately reflect professional practice. To connect the examination content to the knowledge, skills, and abilities that are required for safe and effective practice, a job or practice analysis is typically conducted.[1] The *Standards for Educational and Psychological Testing*[2] state that "…in developing licensure and certification tests, practice analyses or job analyses usually provide the basis for defining the test plan specifications. . ."; (Comment on Standard 4.2, p. 86).

A job analysis or practice analysis can take many different forms. Often, surveys are employed to ask practitioners about the relative importance, frequency, and/or criticality of certain domains or tasks that have been previously defined as being related to the practice of the profession.[3] The content specifications, or blueprint, for a certification examination specify the scope of the domain to be measured and the weighting of the content catego-

This article was externally peer reviewed.
Submitted 8 March 2019; revised 3 June 2019; accepted 7 June 2019.
From The American Board of Family Medicine, Lexington, KY (TRO, MRP, JCP); University of Minnesota Mankato Family Medicine Residency Program, Mankato, MN (KLS); Tidewater Physicians Multispecialty Group, Newport News, VA (JEB).
Funding: none.
Conflict of interest: none declared.
Corresponding author: Thomas R. O'Neill, PhD, 1648 McGrathiana Pkwy, Suite 550, Lexington, KY 40511 (E-mail: toneill@theabfm.org).

ries.[1] This connection between the test content and the practice of the profession provides essential validity evidence about the claims implied by the test scores.[4] Furthermore, periodic reassessment of the degree to which the content specifications mirror clinical practice is essential as the scope of physician practice can change over time.

The American Board of Family Medicine (ABFM) delivers a certification examination that has been used to certify family physicians since 1970. As noted by Norris et al,[5] the ABFM has previously conducted 4 content validity studies of the examination blueprint. In 1982, a task analysis was conducted by researchers from the University of Massachusetts that identified the knowledge, skills, and abilities of practicing family physicians. In 1993, the content validity study included a review of data from the National Ambulatory Medical Care Survey (NAMCS) and physician surveys of clinical experiences. However, due to questionable results derived from a low response rate, another study was conducted in 1999. That study also suffered from a low response rate but provided similar results to the 1993 study. Finally, in 2005, a study using physician surveys of practice content, International Classification of Diseases–Ninth Revision (ICD-9) and Current Procedural Terminology (CPT) codes from physician electronic medical records, and NAMCS data were conducted. Based on the 2005 study, content specifications were revised to use body system categories as shown in Table 1.[5]

The purpose of this study is to review current physician practice and determine the degree to which current practice mirrors the content specifications on the ABFM's Family Medicine Certification Examination (FMCE). More specifically, how well does the FMCE reflect current practice when both frequency and criticality are considered? Several weighting schemas are presented and compared with the current FMCE test specifications.

Table 1. 2006 to 2019 Family Medicine Certification Examinations Content Proportions

A. Organ systems	90%
1. Respiratory	13%
2. Cardiovascular	12%
3. Musculoskeletal	12%
4. Gastrointestinal	7%
5. Special sensory (visual, hearing, etc.)	2%
6. Endocrine	8%
7. Skin	6%
8. Nervous system (brain, spinal cord, peripheral nervous system)	3%
9. Psychogenic (psychological, behavioral, mental health)	7%
10. Reproductive (male)	1%
11. Reproductive (female)	4%
12. Renal/urinary tract	3%
13. Blood/immune system	3%
14. Nonspecific	9%
B. Population-based care and health systems	5%
1. Health policy	
2. Bioterrorism	
3. Legal	
4. Epidemiology	
5. Biostatistics	
6. Evidence-based medicine	
7. Quality improvement	
8. Informatics	
C. Patient-based care and systems	5%
1. Physician-patient interactions	
2. Communication	
3. End of life care	
4. Palliative care	
5. Family issues	
6. Cultural issues	
7. Clinical decision making	
8. Evidence-based medicine	
9. Ethics	

Methods

Data

ABFM FMCE Content Category Specifications

The FMCE content category proportions that have been in place from 2006 through 2019 were based on the 2005 study by Norris et al.[5] These content category proportions were used as the baseline for comparisons.

NAMCS 2012 Data

From the NAMCS 2012 dataset,[6] the ICD-9 codes seen by family physicians were extracted and the frequency of each code was calculated. These data were used in a previous study by Peabody et al,[7] which categorized the ICD-9 codes into the content categories used on the FMCE by a panel of family medicine subject matter experts. The NAMCS data also provided a patient weight for each visit in their sample, which makes it possible to calculate the frequency of each ICD-9 code in the national population. Peabody et al[7] used these weights to make their frequency of ICD-9 codes representative of physician-patient visits nationally.

These nationally representative ICD-9 code frequency weights were also used in this study. To normalize the proportion of the total for each ICD-9 code, the NAMCS-population frequency estimate for each ICD-9 code was divided by the total of the frequencies across all observed ICD-9 codes. This produced a proportion for each ICD-9 code that across all observed ICD-9 codes summed to 1.0.

Criticality Measures for ICD-9 Codes

In the Peabody et al[7] study, the authors also created a criticality index, or Index of Harm scale, in which each ICD-9 code was assigned a criticality value. This assignment was accomplished by asking the subject matter experts on the panel, "How critical is the diagnosis and treatment of this condition?" using a 4-point Likert-type rating scale (Minimally, Moderately, Somewhat, Very). The ratings were then transformed using a Rasch rating scale model[8,9] into interval scale measures that were adjusted for the severity of the individual raters. For the purposes of estimating recommended content category proportions, these measures were then normalized to make the sum of the criticality measures across all the ICD-9 codes equal 100. This was done by calculating the sum of the criticality measures across all ICD-9 codes, and then dividing each measure by that sum and multiplying by 100. For use in this study, the criticality values in the Peabody et al[7] study were divided by 100 to produce an index that ranged from 0 to 1.0 making the range of the scale comparable to the frequency scale.

Schemas for Balancing Frequency and Criticality

Both the frequency of occurrence and the criticality of the relevant tasks are important with regard to deciding how many questions from each blueprint category should be included on the examination. There are methods, such as those described by Spray and Huang,[10] for combining frequency and criticality to produce a set of recommended content category proportions. This study employed a procedure similar to that. When combining weights for frequency and criticality, larger numbers must mean more frequent or critical content and therefore deserve greater representation on the test, while lower numbers must mean the opposite. The ranges of the frequency and criticality scales also must be normalized to have a comparable range. In this study, the authors selected 0 to 100 for convenience.

This study presents 7 different schemas that vary in the weighting of frequency and criticality, and then presents the resulting content category proportion structure for each schema. The schemas will range from frequency-only to criticality-only with 5 incremental changes in the weighting between these 2 extremes. The first schema is frequency-only. The second schema triple weights frequency, but only single weights criticality (3**frequency* + *criticality*). The third schema double weights frequency, but only single weights criticality (2**frequency* + *criticality*). The fourth and middle schema weights frequency and criticality equally. The fifth schema single weights frequency, but double weights criticality (*frequency* + 2**criticality*). The sixth schema single weights frequency, but triple weights criticality (*frequency* + 3**criticality*). The seventh and last schema is criticality-only.

Normalized weights for frequency and criticality were computed separately. When normalized frequency and criticality weights were combined, the resulting weights were renormalized. The renormalized weights were then aggregated within FMCE content category to produce the recommended proportion of the examination that content category should have.

Analysis

Using each of the 7 different schemas, a suggested proportion for each content category on the test was computed. Across the 7 schemas, summary statistics (mean, SD, min, and max) were computed for each of the content categories. In addition, for each of the 7 schemas, scatterplots were created by plotting the sixteen content category proportions of the FMCE against the proportions suggested by the 7 schemas. The correlations associated with these scatterplots were also computed. Furthermore, the correlation between the content category proportions for the frequency only condition and the criticality only condition was also computed.

Results

The relative proportions for each category of the examination that is suggested by the NAMCS frequency data,[6] the Peabody criticality data,[7] and the 5 differently weighted combinations of the 2 are presented in Table 2 and Figure 1. As expected,

Table 2. Content Specifications by Category and Associated Weights

Body System	Current Core	Frequency Only	3F + C	2F + C	F + C	F + 2C	F + 3C	Criticality Only	Mean	SD	Min	Max
Cardiovascular	12	17.6	15.3	14.6	13.0	11.5	10.8	8.5	13.0	1.93	8.5	17.6
Endocrine	8	6.9	6.2	6.0	5.5	5.1	4.9	4.2	5.5	0.56	4.2	6.9
Gastrointestinal	7	5.9	6.8	7.1	7.7	8.3	8.6	9.4	7.7	0.76	5.9	9.4
Hematologic/Immune	3	1.3	1.9	2.1	2.5	2.9	3.1	3.7	2.5	0.51	1.3	3.7
Integumentary	6	5.2	6.2	6.5	7.1	7.8	8.1	9.0	7.1	0.81	5.2	9.0
Musculoskeletal	12	14.6	14.9	14.9	15.1	15.3	15.4	15.6	15.1	0.23	14.6	15.6
Nephrologic	3	2.7	3.3	3.5	3.9	4.2	4.4	5.0	3.9	0.46	2.7	5.0
Neurologic	3	4.5	5.6	6.0	6.7	7.4	7.8	8.9	6.7	0.92	4.5	8.9
Nonspecific	9	7.2	6.9	6.8	6.6	6.4	6.3	6.0	6.6	0.25	6.0	7.2
Psychogenic	7	8.3	7.9	7.8	7.6	7.4	7.3	6.9	7.6	0.25	6.9	8.3
Reproductive-Female	4	3.2	4.1	4.3	4.9	5.4	5.7	6.5	4.9	0.69	3.2	6.5
Reproductive-Male	1	1.1	1.4	1.4	1.6	1.7	1.8	2.1	1.6	0.18	1.1	2.1
Respiratory	13	12.5	11.0	10.5	9.6	8.6	8.1	6.7	9.6	1.23	6.7	12.5
Special Sensory	2	4.1	4.6	4.8	5.2	5.6	5.8	6.4	5.2	0.51	4.1	6.4
Patient-based Systems	5	4.1	3.2	2.9	2.3	1.6	1.3	0.4	2.3	0.81	0.4	4.1
Population-based Care	5	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.00	0.7	0.7
TOTAL	*100*	*99.9*	*100*	*99.9*	*100*	*99.9*	*100.1*	*100*	–	–	–	–

SD, standard deviation.
3F + C = 3*frequency + criticality.
2F + C = 2*frequency + criticality.
F + C = frequency + criticality.
F + 2C = frequency + 2*criticality.
F + 3C = frequency + 3*criticality.

frequency and criticality represent the ends of the spectrum and the weighted combinations of frequency represented incremental steps from pure frequency to pure criticality.

Table 3 and Figure 2 demonstrate the relationship between the 2006 to 2019 content category weights and the weights suggested by the 7 different frequency and criticality conditions. The correlations range from 0.50 to 0.90. The correlation between the pure frequency condition was the highest and the pure criticality was the lowest. Correlation between the only frequency-based content category proportions and the only criticality-based content category proportions was 0.65. The correlation of the current FMCE content category proportions is highest when frequency is weighted more heavily, and it decreases as criticality is given additional weight.

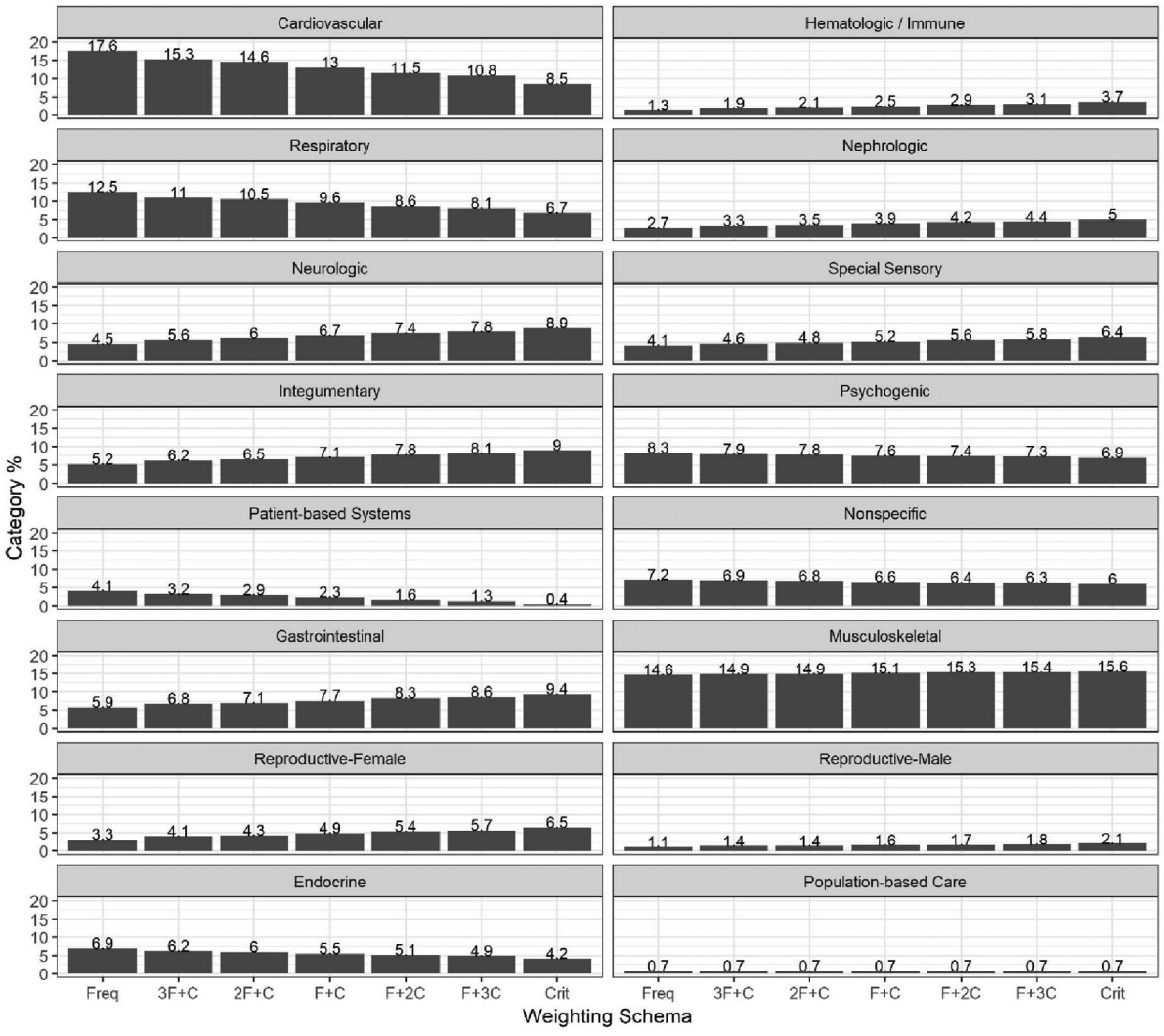
Discussion

The 2006 to 2019 specifications for FMCE content category proportions had the highest correlation (R = 0.90) with the content category proportions that were based on the 2012 NAMCS frequency data. This is not surprising because the current content category proportions were based on the 2004 NAMCS frequency data. This high correlation shows that the NAMCS frequency data across this period of time produced very similar content category proportions. It also suggests that the current FMCE content category proportions continue to be representative of family physician practice in the United States at least with regard to the frequency of what the physicians encounter. It should be noted that this conclusion is limited to the extent that the questions on the examination in each of the categories are in fact representative of the ICD-9 codes assigned to each of the categories.

The 2006 to 2019 FMCE content category proportions had the lowest correlation (R = 0.50) with the content category proportions that were based on the criticality ratings. This is also not surprising because criticality is conceptually different from the frequency of specific kinds of patient-physician encounters.

Both criticality and frequency of occurrence are important considerations, but are they equally important? The relative weight of the frequency and

Figure 1. Stability of content category proportions (from least to most) across weighting schemas.



criticality of each ICD-9 code need not necessarily be a strict 1-to-1 relationship. For instance, perhaps a rarely seen condition with a high criticality should be given more weight than a frequently seen yet low criticality condition. The issue in choosing a relative weighting schema for frequency and criticality is really about what is the test intended to measure. What is the construct? This discrepancy points out the tension between these covering a wide breath of family medicine and the aspects that are critical to diagnose and treat correctly the first time.

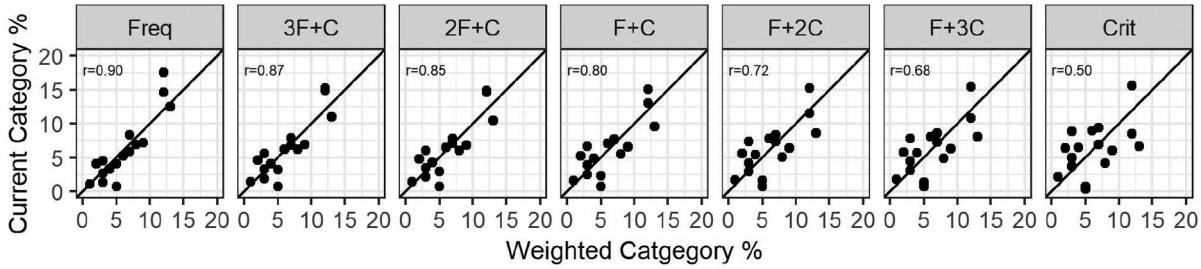
In addition, both Patient-Based Systems and Population-Based Care were added as 5% of the examination blueprint without considering their frequency as they are hard to reconcile in the NAMCS data. So the question of category propor-

Table 3. Correlation of the Family Medicine Certification Examination Content Category Proportions with Proportions Suggested Using Other Weighting Schemas

Weighting Schema	Correlation	N
Frequency only	0.90	16
3*Frequency + Criticality	0.87	16
2*Frequency + Criticality	0.85	16
Frequency + Criticality	0.80	16
Frequency + 2*Criticality	0.72	16
Frequency + 3*Criticality	0.68	16
Criticality only	0.50	16

Note: The correlation between the frequency condition and the criticality condition is 0.65, suggesting that they are somewhat similar but not identical.

Figure 2. Comparison of the FMCE content proportions with the proportions suggested by other weighting schemas. FMCE, Family Medicine Certification Examination.



tions is really a policy decision that will not have an empirical answer; however, hopefully this study can assist the policy makers in their decisions about the examination content.

Limitations

Although the criticality ratings were made in 2017, the frequency data are from 2012, making them 7 years old. The 2012 frequency results are similar to the 2005 study, which suggest that the types of patient interactions were not terribly different, but it would be desirable to be able to get data that is available in a timelier manner. In the future, it would be preferable for ABFM to survey a representative sample of their diplomates. This would limit the observations to what is being seen by board certified family physicians, which seems appropriate for setting the specifications for a board certification examination. In addition, the ICD-9 code often did not lend itself to being neatly classified by the ABFM certification examination's content category classification system, which is based on body systems.

Conclusion

The results of this study support the continued use of the current FMCE content specifications as being representative of current family medicine practice; however, small adjustments might be warranted to permit the criticality of the topics to be better represented. In the end, the test plan specifications are policy decisions defining what a test is intended to measure. Empirical evidence can be used to support the reasonableness of that policy decision, but there is not an empirical solution to what the specifications should be.

These results should not be interpreted to mean that a different content category classification schema would not work for the FMCE. Other classifications systems could include grouping content using a "presenting problem perspective," an acuity perspective, a patient age perspective, etc. Describing the content on the examination is important to those who are preparing to take it and can shape how they prepare for it, so changes to the content category classification schema that is actually used to craft the examination should be articulated clearly to all examinees. Alternative classification systems that are not used to craft the examination can still be used as an additional way to parse feedback to the examinees, but there is typically no promise that those alternative schema categories will be seen in the same proportion on different forms of the examination.

Periodic assessment of current medical practice and using that information to create and inform policy on the examination specifications is necessary to assure that the examination does not drift from the parameters of medical practice.

To see this article online, please go to: http://jabfm.org/content/32/6/876.full.

References

1. Raymond MR, Neustel S. Determining test content of credentialing examinations. In: Downing S, Haladyna T, eds. Handbook of test development. Mahwah, NJ: Erlbaum; 2006;181–224.
2. AERA, APA, NCME. Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 2014.
3. Raymond MR. A practical guide to practice analysis for credentialing examinations. Educ Meas Issues Pract 2002;21:25–37.
4. Lane S, Raymond MR, Haladyna TM. Test development process. In: Handbook of test development. New York, NY: Routledge; 2016;3–18.
5. Norris TE, Rovinelli RJ, Puffer JC, Rinaldo J, Price DW. From specialty-based to practice-based: a new

blueprint for the American Board of Family Medicine cognitive examination. J Am Board Fam Pract 2005;18(6):546–54.
6. National Center for Health Statistics. National center for health statistics ambulatory health care data. 2012. Available from: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NAMCS/.
7. Peabody MR, O'Neill TR, Stelter KL, Puffer JC. Frequency and criticality of diagnoses in family medicine practices: from the National Ambulatory Medical Care Survey (NAMCS). J Am Board Fam Med 2018;31(1):126–38.
8. Andrich D. A rating formulation for ordered response categories. Psychometrika 1978;43:561–73.
9. Wright BD, Masters GN. Rating scale analysis. Chicago, IL: MESA Press; 1982.
10. Spray JA, Huang CY. Obtaining test blueprint weights from job analysis surveys. J Educ Meas 2000; 37:187–201.