Criterion-Referenced Examinations: Implications for the Reporting and Interpretation of Examination Results

Kenneth Royal, PhD, and James C. Puffer, MD

The purpose of the American Board of Family Medicine (ABFM) certification/maintenance of certification examination is to measure the basic knowledge necessary to deliver high-quality care to patients and their families. More than 25 years ago, the ABFM became the first American Board of Medical Specialties board to introduce criterion-based methodology to establish the passing threshold for its examination. A criterion-referenced examination is one in which a particular score is required to pass, and the performance of those taking the examination is of no consequence in determining who passes or fails. In other words, all candidates taking the examination could theoretically pass if they met or exceeded the criterion-referenced passing score. Furthermore, the examination is equated across forms and administrations, meaning candidates are not advantaged or disadvantaged by having received a particular version of the examination or by taking it at a particular time of the year.

It should be apparent, therefore, that the ABFM is not interested in comparing the performance of one candidate with another, but rather comparing a candidate's performance against the criterion-based passing threshold. ABFM's ability to do so became more precise in 2006 when it moved to a new psychometric model, Item Response Theory, to develop and score the examination. Among its many advantages over the Classical Test Theory model that had been employed for more than 35 years, Item Response Theory provides greater discrimination and precision around the passing threshold. However, it also provides less useful information for those who score very well or very poorly, and that is one of the major reasons why the ABFM recently has discontinued the use of percentile ranks associated with a candidate's score. Reporting percentile ranks can be problematic and potentially misleading for examinees, and the ABFM would like to demonstrate why that is so.

Because candidates who apply for the examination consist of both recently trained residents seeking certification for the first time as well as seasoned family physicians seeking to maintain their certification, the cohort of family physicians who sit for the examination each year is quite diverse. The demographic characteristics, experience level, geographic location, and even scope of practice of the physicians in each sample vary considerably. This was particularly true for the cohorts that took the examination in 2010, 2011, and 2012.

Before 2005, the ABFM granted certification for 7-year periods. Beginning in 2005, a policy change was implemented within the Maintenance of Certification for Family Physicians (MC-FP) program that created the possibility for family physicians to earn a 3-year extension of their certificate, thereby extending the period of time between examinations to 10 years. As a result of this policy change, the ABFM experienced a 3-year period in which the number of family physicians seeking to maintain their certification was very low. However, the number of family physicians who previously had failed and were attempting to recertify was disproportionately high. This phenomenon is best demonstrated by comparing the 2009 and 2010 examination cohorts.

In Table 1, percentile ranks are reported for both the 2009 and 2010 MC-FP exams. The passing standard for the examination in both years was 390, with a reported scaled score range of 200 to 800. Because the cohorts of initial certifiers (primarily residents) in 2009 and 2010 were relatively stable, the percentile rank did not change much from 2009 to 2010 (approximately 2 percentile

Conflict of interest: The authors are from the ABFM.

Table 1. Percentile Rank Comparisons for Initial Certifiers and Recertifiers for 2009 and 2010

Scaled Score	Initial Certification Candidates (n)		MC-FP Candidates (n)	
	2009	2010	2009	2010
300	2	3	3	10
310	3	3	4	11
320	4	4	5	13
330	4	5	6	15
340	6	7	7	17
350	7	8	8	19
360	9	10	9	22
370	11	12	11	25
380	13	15	13	28
390*	16	18	15	31
400	19	21	17	34
410	22	24	19	37
420	26	28	22	41
430	30	32	24	44
440	34	36	27	48
450	38	40	30	51
460	43	45	34	55
470	47	49	37	59
480	52	54	40	62
490	56	58	44	65
500	61	63	47	69
510	65	67	51	72
520	69	71	54	75
530	73	75	58	77
540	77	78	61	80
550	80	81	65	82
560	83	84	68	85
570	86	87	71	87
580	89	89	74	89
590	91	91	77	90
600	92	93	79	92
610	94	94	82	93
620	95	95	84	94
630	96	96	86	95
640	97	97	88	96
650	98	98	90	97

MC-FP, Maintenance of Certification for Family Physicians.

points) for these candidates. However, for those attempting to maintain their certification, a scaled score of 390 in 2009 meant one was in the 15th percentile. In 2010, however, that same scaled score meant one was in the 31st percentile. One will note other significant differences when scanning Table 1 as well.

It is interesting that many examinees can recall their percentile ranking but cannot recall their scaled score. It is easy to understand why some examinees may be interested in learning how well they performed relative to their peers. Yet, from the example described earlier, it is evident that percentile rankings may be misleading for both examinees and the general public. When the ranking portrays the examinee as being more knowledgeable than he or she truly is, it inflates and misrepresents one's perceived ability and misleads the public. For example, consider an MC-FP candidate in 2010 that scored a 450 on the examination and wants to compare the ranking with other candidates. This examinee would rank in the 51st percentile among his or her MC-FP peers, but only in the 40th percentile when compared with candidates seeking initial certification.

The practice of reporting percentile rankings has the potential to introduce other undesirable elements into the score reporting process as well. For example, the very nature of reporting percentile ranks will no doubt mean some people will be pleased with their ranking, whereas others will not. After all, those at the top of the scale will certainly feel great about themselves knowing they outperformed the vast majority of their peers on a national examination. However, for those unfortunate examinees who happened to fail the examination it can be rather embarrassing to realize that, for example, 96% of one's peers performed better than he or she did. When an examination is criterion-referenced, the only thing that really matters is one's performance relative to the minimum passing standard. After all, someone who scores a 500 on the MC-FP examination is not "more certified" than someone who passed with a score of 400. The ABFM contends that through reporting scores properly and directing examinees toward the appropriate criteria for making meaningful inferences, it can be more responsible with score reporting while concurrently preserving the dignity of those who inevitably fail.