The IIP Examination: an Analysis of Group Performance 2009–2011

Ben Babcock · Paul Nagy

Published online: 9 January 2013
© Society for Imaging Informatics in Medicine 2013

Abstract This report summarizes the performance of the Imaging Informatics Professional (IIP) examination from 2009 to 2011 (six exam administrations). Results show that the IIP exam is a reliable measuring instrument that is functioning well to consistently classify candidates as passing or failing. An analysis of the section scores revealed that content in the Image Management, Systems Management, and Clinical Engineering sections of the exam were somewhat more difficult than the content in the other sections. The authors discuss how future candidates may use this information to help hone their study strategies. By all indications, the IIP examination appears to be statistically functioning as a high-quality certification measuring instrument.

Keywords ABII · IIP · Certification · Examination · Statistics

Background

In 2007, the Society for Imaging Informatics in Medicine (SIIM®) and the American Registry of Radiologic Technologists (ARRT®) formed the American Board of Imaging Informatics (ABII®). The purpose of ABII was to provide a certification mechanism for radiologic technologists and information technology personnel that work with imaging informatics. To this end, ABII created the Imaging Informatics Professional (IIP®) examination as a major step in the certification process [1]. This report reviews the

statistical performance of the IIP exam for the three calendar years, or six administrations, from 2009 to 2011. This study excludes the first two exam administrations because often, the initial testing groups for a new certification exam are not representative of the true exam population and can bias results. Candidates waiting for the exam to first become available often have years of experience.

It is important to periodically review the statistical properties of any exam. First, such review is vital to the creators of the exam, because the statistical properties provide a feedback mechanism to maintain or improve the quality of test construction. Second, a review of a test's statistical properties is useful to candidates before taking an exam. This informs candidates of how difficult the exam has been on average to other candidates and which sections of the exam proved most challenging.

Methods

The first question that many people ask about an exam is “what percentage of people pass the exam?” To this end, the authors first examined basic passing statistics. Information about the overall test scores, however, gives much more information than simply looking at passing statistics. The authors also conducted a basic descriptive analysis using these total exam scaled scores.

It is important to know how ABII reports the scores of the IIP examination so that one may correctly interpret scores. There are three types of scores that enter the score reporting conversation: raw scores, percentage correct scores, and scaled scores. Raw scores are simply the number of questions that a person answers correctly.¹ Percentage correct scores are simply the raw scores divided by the total number

B. Babcock (✉)
The American Registry of Radiologic Technologists,
1255 Northland Drive,
St. Paul, MN 55120, USA
e-mail: ben.babcock@arrt.org

P. Nagy
Johns Hopkins University,
Baltimore, MD, USA

¹ ABII scores all of its test questions correct/incorrect, or 1/0. There is no partial credit for any of the questions on the exam.

of items on the test, multiplied by 100. The ABII does *not* report raw scores. The ABII also does *not* report percentage correct scores. The reporting system for IIP test results is an industry standard methodology called scaled scoring [2]. No two exam administrations have the exact same questions, which can translate into varying exam difficulty from administration to administration. If ABII required a strict raw score to pass the exam for every single administration, then examinees who took a more difficult exam form would be disadvantaged compared to individuals who took easier exams. ABII's psychometrics staff statistically adjusts every score using well-researched methods so that no candidate is at a disadvantage based on the particular exam that he or she takes. The statistical process is known as equating [3]. The number of correct required for passing varies from administration to administration based on the equating results. Scaled scores are always integers. The scaled score required to pass is always 75, with 1 as a minimum and 99 as a maximum possible score on the scale. Scales similar to this are common in the testing industry [4, Ch. 2].

As an example of calculating a scaled score from a raw score, suppose that the minimum score required to pass a given exam form (i.e., raw cutscore) was 83 correct out of the 130 scored questions on the exam. ABII determines the scale for the scores by doing a simple linear transformation of scores. The slope b for the line is

$$\begin{aligned} b &= (\text{Max Score}_{\text{Scaled}} - \text{Cut Score}_{\text{Scaled}}) \\ &\quad / (\text{Max Score}_{\text{Raw}} - \text{Cut Score}_{\text{Raw}}) \\ &= (99.49 - 74.5) / (130 - 83) = 0.5317 \end{aligned} \quad (1)$$

Note that the scaled cut score used in this equation is 74.5, which rounds up to 75. The intercept a for the linear transformation equation is

$$\begin{aligned} a &= \text{Cut Score}_{\text{Scaled}} - b \times \text{Cut Score}_{\text{Raw}} \\ &= 74.5 - 0.5317 \times 83 = 30.3689 \end{aligned} \quad (2)$$

Suppose that a person taking this exam obtained a raw score of 87 out of 130. The person's total scaled score would be $30.3689 + 0.5317 \times 87 = 76.6268 = 77$ (after rounding). The scaled score is 75 or greater, so the person passed the exam. If a person's raw score was 81, the person's scaled score would be $30.3689 + 0.5317 \times 81 = 73.4366 = 73$ after rounding. The scaled score for this person is less than 75, so the person did not pass the exam. Note that neither one of the scaled scores was the same as the raw score nor the percentage correct for the test taker. The above type of linear transformation does not affect the relative spacing of scores within a test form [5], and it forces the passing standard to be the same score of 75 across all test forms.

The reader should note that the maximum possible raw score was 130 even though there are 150 questions on the IIP exam. The additional 20 questions are called pilot questions. Pilot questions are tryout questions, so they do not count toward people's scores. The pilot testing process helps ABII evaluate new questions, often about new content, to ensure that the all questions function properly before the questions affect examinee scores.

ABII also reports scaled scores for the 10 individual sections of the IIP examination. The authors conducted a brief analysis of these scores to determine which sections were the most difficult. Before looking at the section scaled score statistics, it is very important to discuss how section scaled scores work. ABII scales the section scores from 1 to 9.9, with 7.5 being "on pace" in that particular section to pass.² Section scores are much less reliable statistically than total scores, because individual sections are much shorter than the whole exam and, therefore, have scores with lower reliabilities [6, chapter 6]. The reader should always use caution when interpreting section scores. The ABII provides section scaled scores for guidance purposes only; one cannot add up or average section scaled scores to arrive at the reported total score. ABII staff calculates total scaled scores directly from total raw scores and then uses that information to derive the section scores. ABII staff calculates section scaled scores after the total scaled scores rather than use the sections to determine the total scaled score.

When evaluating the statistical functioning of a credentialing examination, it is important to examine how reliable the examination is, as well as how consistently the examination classifies examinees. This report contains both a reliability analysis and a decision consistency analysis for the IIP examination. Reliability is the statistical property of repeatability of exam scores. Reliability in this context refers to what proportion of the measurements are reliable "signal" [6 Ch. 5], as opposed to "noise." Reliability coefficients range from zero to one, with higher numbers indicating higher "signal detection" in the measurement and lower numbers indicating mostly "noise" in the measurement. This report used the Kuder–Richardson Formula 20 [7, Ch. 7] as the reliability estimate.

If there is one statistic that is even more important to credentialing examinations than reliability, it is decision consistency. Making good and consistent decisions about who fails and who passes the examination is vitally important to ensuring the integrity of a certification examination. Classification consistency refers to how often a person taking an examination on multiple occasions would receive the same pass/fail decision. This report calculates decision

² Note that 7.5 is 75 divided by 10. The ABII places the section scaled scores on a 1 to 9.9 scale instead of a 1 to 99 scale to emphasize that the scores are not as important as the overall scaled score.

consistency using Subkoviak's [8] method, which is relatively common when examining consistency in credentialing examinations.

It is also important, however, to take into account statistical chance when examining pass/fail decisions. Suppose that a person had a room full of IIP candidates, all of whom qualified and studied for the IIP exam. A person randomly assigning pass/fail decisions to 86.8 % (the first-time passing percentage from Table 1) of a group of people of whom 86.8 % are indeed qualified would make consistent pass/fail decisions approximately 77 % of the time ($[0.868]^2 + [1 - 0.868]^2$; 8). This chance correct consistency is a well-documented phenomenon, such as when diagnosing the presence or absence of a disease. In order to adjust for this statistical chance, one must calculate the statistic kappa. Kappa shows proportion of individuals consistently classified beyond chance consistency. The equation for kappa is

$$k = \frac{p_0 - p_c}{1 - p_c}, \quad (3)$$

where p_0 is the overall consistency of certified/not certified classifications and p_c is the proportion of consistent classifications that would be expected by chance [9]. This report calculated kappa in order to determine how well the IIP examination classified individuals above and beyond chance consistency.

Results

Basic Statistics for Passing Rates and Total Scaled Scores

Table 1 contains passing statistics for the IIP certification exam. 86.8 % of first-time test takers passed the examination. The passing percentage for repeat test takers, however, is much lower at around 50 %. This result is typical across numerous certification programs, because the most prepared examinees tend to pass on the first attempt.

Table 2 contains some total test scaled score statistics for first-time test takers. The mean, or arithmetic average, is well above the minimum passing score of 75. This is not surprising considering that most first-time examinees pass the IIP exam. The first quartile, the score point below which 75 % of scores occur, is 79, which also indicates that most people exceed the passing score by several scaled points.

Table 1 IIP passing statistics

Group	Number taking	Passing %
First-time takers	432	86.8
Repeat takers	36	52.8

Table 2 Total test scaled score descriptive statistics, first-time test takers only

Mean	Standard deviation	First quartile	Median	Third quartile
82.9	7.3	79	84	88

A scaled score of 75 is passing

In addition to statistical tables, histograms give a good picture of the shape of the distribution of test scores. Figure 1 is a histogram of total test scaled scores. The distribution of scores has a "heavier tail" at the low end of the score continuum than at the high end. This is called negative skew; this type of skew is quite common in certification testing programs [10]. Scores varied substantially across examinees, but the largest block of scores was between the scaled scores of 80 and 90.

Section Scaled Score Statistics

In order to visualize section scaled score data, Fig. 2 gives side-by-side boxplots of the section score data. For each section, the thick line inside the box represents the median (the middle score), the top and bottom of the box are the third and first quartiles, respectively, the dotted lines (whiskers) represent the range where one could reasonably expect most scores to be, and the dots are relative outliers. The box portion of each boxplot represents the middle half of the data.

The easiest sections, based on the high median scores, were Training and Education, Project Management, and Communications. The most difficult sections were Image Management, Systems Management, and Clinical Engineering.

Fig. 1 Histogram of total test scaled scores, first-time test takers only

Fig. 2 Boxplots of section scaled scores, first-time test takers only with a scaled score of 7.5 reference line

Information Technology and Medical Informatics were also relatively difficult.

Reliability and Classification Consistency

Table 3 contains reliability estimate statistics for the IIP examination. The mean reliability was 0.89. Reliability estimates ranged from 0.84 to 0.92 across the six administrations of the IIP examination. These reliability estimates indicate that the IIP examination was a reliable measuring instrument across 3 years of reporting data.

The classification consistency as reported in Table 3 is 0.94. This indicates that the IIP examination had consistent classifications for 94 % of its examinees, which is quite high. The kappa for the IIP examination from Table 3 is 0.74, which indicates that the IIP examination made classification decisions that were 74 % better than chance. This is quite good for a certification exam.

Discussion

The statistical analysis above indicates that the IIP examination is functioning quite well. The first-time pass rate of 86.8 % indicates that the candidates who take the IIP exam are prepared and know a great deal about imaging informatics going into the exam, but it is still a challenge to pass the

exam. The total test scaled score statistics highlight that examinees obtain a wide range of possible scores, which is a desirable property of examinations. The reliability and decision consistency statistics indicate that the IIP exam is functioning as desired.

An individual preparing for the IIP exam could look at the section scaled score analysis and draw some conclusions about study strategies. First, the Image Management section is both the exam's longest section and one of its most difficult.³ Studying this material gives test takers the greatest potential on average for improving their score. The Information Technology section is also relatively long and relatively difficult. When forming a study strategy, however, it is important to look over materials for all of the content as outlined in the IIP exam's test content outline (<https://www.abii.org/docs/Exam-Content-Outline.pdf>). No single study strategy is best for everyone, so individuals must judge the difficulty of the sections, the percentage of items, and their own knowledge when deciding where to study the most.

Conclusion

The IIP exam has matured into a stable psychometric test to evaluate a candidate's knowledge in imaging informatics. The exam provides a significant hurdle for certification so that those passing the exam are well qualified. The pass rate is in alignment with other professional certifications. The IIP exam has some sections that are slightly more difficult than others.

³ Image Management was significantly more difficult than the first five sections according to the results of a Hierarchical Linear Model. The differences between Image Management and the other sections were non-significant.

Table 3 Reliability and classification consistency indices for total scores, first-time test takers only

Mean KR-20 reliability	Min. KR-20	Max. KR-20	Classification consistency	Chance consistency	Kappa
0.89	0.84	0.92	0.94	0.77	0.74

Acknowledgments The authors would like to thank the ABII Board of Trustees, the ABII staff, the IIP examination committee, the IIP item writers, and the hundreds of CIIP® diplomats for making the IIP examination the quality measuring instrument it is today.

References

1. American Board of Imaging Informatics: <https://www.abii.org/>, Accessed 15 March 2012
2. Thissen, D, Wainer, H: Test Scoring. Mahwah, NJ: Erlbaum, 2001
3. Babcock B, Albano A, Raymond R: Nominal weights mean equating: A method for very small samples, Ed and Psych Measurement, 2012
4. Kolen, M J, Brennan, R L: Test Equating, Scaling, and Linking: Methods and Practices, 2nd ed. New York: Springer, 2004
5. Howell, D C: Statistical Methods for Psychology, 6th Ed, Ch. 2. Belmont, CA: Thompson Wadsworth, 2007
6. Furr, R M, Bacharch, V R: Psychometrics: An Introduction. Thousand Oaks, CA: Sage, 2008
7. Crocker, L, Algina, J: Introduction to Classical and Modern Test Theory. Belmont, CA: Wadsworth, 1986
8. Subkoviak, M J: Estimating reliability from a single administration of a mastery test. J of Ed Measurement 13: 265–276, 1976
9. Cohen, J: A coefficient for agreement for nominal scales. Ed and Psych Measurement 20: 37–46, 1960
10. Micceri, T: The unicorn, the normal curve, and other improbable creatures. Psych Bul 105:156–166, 1989