TECHNICAL APPENDIX TO THE ARRT'S PRIMARY EXAM RESULTS

2012

Introduction

This report summarizes the psychometric characteristics of ARRT's examinations in Radiography (RAD), Nuclear Medicine Technology (NMT) and Radiation Therapy (THR) for the year 2012. This report is a companion document to the *Primary Exam Results* report.

The first section of this report contains information about the duration of time that candidates used to complete their examinations. The second section provides descriptive statistics of total exam scores, both raw and scaled, and information about how ARRT converts raw scores to scaled scores. The third section of this report presents descriptive statistics for the exams' section scores, including correlations and reliability estimates. Section four provides more detail about the reliability of the overall exam scores, with a discussion of coefficient α and the standard error of measurement. The final section of the report addresses decision consistency, which quantifies the reproducibility of the certification decisions that ARRT makes based on its examinations.

Information about Exam Durations

Most certification exam administrators, including ARRT, do not intend to have exam administration time be a heavily influential factor for examinees. Practical limitations, however, make it necessary to establish exam time limits. For all three primary examinations, candidates may take up to 210 minutes (3.5 hours) to answer 220 items (questions). The intention of the time limit is to have the exam begin and end in a reasonable amount of time, while also ensuring that knowledgeable candidates have sufficient time to complete the exam assuming that they remain focused. It is ARRT's intention that, although its exams are time limited, its exams are not speeded exams.

This section presents information on the amount of time that examinees used to take the three exams described in this report. Some sources (e.g., Nunnally, 1978) specify that an exam is unspeeded when at least 90% of examinees complete the exam within the allotted time. If results show that more than 10% of examinees require all 210 minutes, ARRT would re-evaluate existing time limits.

Table 1 contains a summary of the amount of exam time spent by first-time examinees in 2012. These and all other statistics reflect only first-time ARRT examinees. None of the statistics include state candidates, people retaking the exam after failing the initial attempt, or people taking the Americans with Disabilities Act special forms. This table indicates that THR candidates spent more time than their counterparts in NMT and RAD. THR had the highest mean (average) time and had the smallest standard deviation, which indicates less variation in the amount of time spent taking the exam.

Table 1. Descriptive Statistics of Examinees' Time Spent on Examination (in Minutes)

Discipline	Number of Candidates	Minimum Time	Maximum Time	Mean Time	Standard Deviation
RAD	*12,333	45	210	143.60	37.49
NMT	+460	58	210	142.97	38.84
THR	+894	68	210	167.56	33.91

*Excludes 5 ADA candidates. +Excludes 1 ADA candidate.

Table 2 divides the candidates into nine groups according to the amount of time for the cumulative group to complete the exam. Using RAD as an example, 10% of all candidates completed the exam in 95 minutes or less, and 20% completed it in 108 minutes or less. Continuing on the row, Table 2 shows that 90% of RAD candidates completed the exam in 198 minutes or less. Overall, most candidates completed their examinations within the established time limits. For all three disciplines, 90% or more of the candidates completed the exam in less than the allotted 210 minutes. These exams do not appear to be speeded under the 90% or more completion criterion.

Table 2. Number of Minutes Required to Complete Exams by Percentiles

Discipline	Cumulative Percentage of Candidates Completing the Exam								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
RAD	95	108	120	131	142	154	167	181	198
NMT	91	106	117	129	140	155	169	183	200
THR	118	135	148	162	172	183	193	203	209

Descriptive Statistics for Total Examination Scores

Table 3 contains descriptive statistics for the raw scores (number correct), which are the basis for numerous other calculations in this report. Although each of the primary exams has 220 items, the total score consists of only 200 items. The additional 20 items are unscored "pilot" items.

Table 3. Descriptive Statistics of Raw Scores

Discipline	Minimum	Maximum	Mean	Standard Deviation
RAD	48	198	159.96	19.14
NMT	73	192	151.82	21.35
THR	83	197	157.30	16.72

ARRT uses scaled scores to report exam results. Total scaled scores range from 1 to 99, and a candidate must achieve a total scaled score of 75 to pass an examination. Table 4 contains descriptive statistics for the total scaled scores, which are from the *Primary Exam Results* report. The main advantage of scaled scores is that they facilitate a meaningful comparison of scores across forms and years.

Table 4. Descriptive Statistics of Scaled Scores

Discipline	Minimum	Maximum	Mean	Standard Deviation
RAD	46	99	85.28	6.80
NMT	59	97	84.00	6.87
THR	57	98	83.49	6.27

In order to convert raw scores to scaled scores, ARRT must determine the difficulty of an exam form. Each exam consists of items that were used on previous exams. ARRT uses the Rasch model, also called the one-parameter logistic IRT model, to track the difficulty levels of individual exam items and, consequently, whole exam forms. Each item has a Rasch difficulty statistic that indicates the probability of an examinee giving a correct answer.

ARRT determines the difficulty of an exam form by calculating the sum of the probabilities of correct answers at the cutpoint. Comparisons with the difficulties of previous forms determine the relative difficulty level of the new form. If the new form is easier, the cut score for the new form will be greater by an appropriate number of questions. If the new form is more difficult, then the cut score will be lower by some appropriate number of questions.

After determining the raw passing score, ARRT calculates equations to convert the raw scores to scaled scores such that the scaled scores range from 1 to 99 with a passing score of 75. As a hypothetical example, assume that the raw passing score is 130 out of 200. The conversion equation requires two scaling coefficients: the slope (a) and the intercept (b). The calculations of a and b involve four values: the maximum scaled score (99.49), the scaled cut score (74.50), the maximum raw score (200) and the raw cut score (130).

$$a = (99.49 - 74.50) / (200 - 130)$$
$$a = .357$$

$$b = 74.50 - (a \times 130)$$
$$b = 74.50 - (.357 \times 130)$$
$$b = 28.09$$

For this hypothetical form, the scaling coefficients would be $a = .357$ and $b = 28.09$. ARRT would use these scaling coefficients to convert the raw scores to scaled scores. If a candidate achieved a raw score of 131 (one point above passing), then the scaled score would be

$$\text{scaled score} = (\text{raw score} \times .357) + 28.09 = (131 \times .357) + 28.09 = 74.857,$$

which rounds up to 75, a passing scaled score. For this example, raw scores of 130 and 131 round up to a passing scaled score of 75. Raw scores of 128 and 129, however, round down to a scaled score 74, which is a failing score.

Table 5 contains the pass percentages for ARRT's primary examinations. This information is also in the *Primary Exam Results* report, but is repeated here because of its importance.

Table 5. Pass Percentages for First-Time Candidates

Discipline	Pass Percentage
RAD	93.03
NMT	90.22
THR	91.57

Descriptive Statistics for Section Scores

In addition to the total scaled score, ARRT reports individual section scores that correspond to the major content areas as outlined in the content specifications of each exam. The primary purpose of the section scores is to provide general information to examinees regarding their strengths and weaknesses in particular content categories. ARRT reports section scores on a scale from 0.1 to 9.9 in one-tenth point intervals. The *Primary Exam Results* report contains descriptive statistics for the scaled section scores.

Section scores are useful to the extent that: (a) the scores are reliable and (b) the sections measure knowledge and skills that are independent of each other. For these reasons, Tables 6 through 8 contain additional descriptive statistics about ARRT's section scores. These include the correlations among the section scores as well as the number of items in each section, raw score means, and standard deviations. In addition, the tables contain a reliability estimate (Cronbach's α) for each section. Sections with more items generally have more reliable scores in the same way that longer examinations generally have more reliable scores. Page 6 discusses reliability in more detail.

The correlations among the section scores provide a measure of their distinctness. In theory, correlations can range from –1.00 (perfect inverse linear relationship) to +1.00 (perfect positive linear relationship). Section scores on an exam are usually positively correlated, because candidates who perform well on one section typically perform well on others. In Tables 6 through 8, the section score correlations above the diagonal are the observed (uncorrected) correlations, and the correlations below the diagonal are correlations corrected for unreliability. The corrected correlations take into account the unreliability of the section scores and give a sense of the magnitude of the correlations under the condition of perfect reliability. For Radiography, the observed correlations ranged from 0.50 to 0.69. After correction, the correlations ranged from 0.77 to 0.94. The high correlations after correction indicate a high degree of common variance among the section scores.

Tables 6-8. Section Score Statistics for RAD, NMT, and THR in 2012

Table 6. RAD Section Score Correlation Matrix and Statistics

Content Area	Rad. Prot.	Equip. Op.	Image Prod.	Rad. Procs.	Patient Care
Rad. Prot.		0.64	0.69	0.64	0.52
Equip. Op.	*0.90		0.68	0.63	0.50
Image Prod.	*0.91	*0.94		0.66	0.55
Rad. Procs.	*0.83	*0.87	*0.84		0.55
Patient Care	*0.78	*0.77	*0.79	*0.78	
Statistic					
No. of Items	45	22	45	58	30
Mean	36.20	17.06	35.13	47.42	24.15
Std. Dev.	4.78	3.16	5.48	6.14	3.32
Reliability	0.73	0.67	0.78	0.80	0.62

Table 7. NMT Section Score Correlation Matrix and Statistics

Content Area	Rad. Prot.	Radionuclides	Instrumentation	Dx. Procs.	Patient Care
Rad. Prot.		0.57	0.59	0.61	0.33
Radionuclides	*0.96		0.69	0.73	0.44
Instrumentation	*0.91	*0.96		0.76	0.49
Dx. Procs.	*0.89	*0.96	*0.92		0.55
Patient Care	*0.63	*0.77	*0.78	*0.82	
Statistic					
No. of Items	20	22	40	100	18
Mean	15.62	16.23	29.48	76.21	14.27
Std. Dev.	2.51	3.14	5.34	11.36	2.25
Reliability	0.54	0.65	0.78	0.88	0.51

Table 8. THR Section Score Correlation Matrix and Statistics

Content Area	Rad. Prot.	Clin. Concepts	Trt. Planning	Trt. Delivery	Patient Care
Rad. Prot.		0.58	0.65	0.47	0.46
Clin. Concepts	*0.89		0.66	0.46	0.49
Trt. Planning	*0.95	*0.87		0.58	0.46
Trt. Delivery	*0.85	*0.75	*0.91		0.33
Patient Care	*0.83	*0.80	*0.72	*0.63	
Statistic					
No. of Items	35	55	55	25	30
Mean	27.52	43.02	41.90	19.99	24.87
Std. Dev.	3.56	5.46	6.27	2.61	2.82
Reliability	0.59	0.73	0.79	0.51	0.52

*Correlation corrected for unreliability

When interpreting the correlations in Tables 6 through 8, it is important to consider the reliability of each section score. Sections with low reliability will have low correlations with other subscales. This is why the report provides the corrected correlations. A low reliability coefficient for a section also indicates that a candidate's score for that section is only an approximation of his or her true level of knowledge. For this reason, ARRT cautions students and program directors not to over-interpret small score differences among section scores. The limited reliability of section scores is the primary reason that ARRT bases its pass/fail decisions on total scores. Total scores are sufficiently reliable to make pass/fail decisions; section scores do not have sufficient reliability to make those decisions.

Reliability of Exam Scores

Reliability refers to the repeatability and consistency of exam scores. An examinee who takes one form of an exam on one occasion and a second parallel form on another occasion should earn similar scores if the exam scores are reliable and the examinee has not changed in the time between the exam administrations (i.e., learned new material). Major differences should occur only if there is true change in the examinee's knowledge or if the exam is unreliable.

Reliability also describes how well candidates' observed scores on an exam approximate their "true" scores. An examinee's true score is the mean of an examinee's observed scores from a very large number of examinations. True score is theoretical and not observable in practice.

Reliability coefficients are estimates of the reliability of exam scores. Reliability coefficients typically range from zero to one, with values near one indicating high consistency and those near zero indicating little or no consistency. In this report, Cronbach's coefficient α is the reliability estimate of choice. Cronbach's α, which requires only one exam administration, is an estimate of the reliability of a group's exam scores. Although it is never possible to determine the exact amount of error in one specific person's score, the standard error of measurement (SEM) describes the expected variation of each examinee's observed score around his or her true score.

Coefficient Alpha

The equation for Cronbach's coefficient α is

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{I} \hat{\sigma}_i^2}{\hat{\sigma}_X^2}\right), \qquad (1)$$

where k is the number of items,
I is the total number of items,
X is a set of exam scores,
$\hat{\sigma}_i^2$ is the variance on an individual item i, and
$\hat{\sigma}_X^2$ is the total exam variance.

Table 9 contains the reliability estimates for RAD, NMT, and THR in 2012. Recalling that reliability coefficients range from 0.0 to 1.0, one can see that the reliability estimates for the three exams are quite high at 0.90 or greater. These high reliability estimates mean that observed scores for these exams correspond quite closely to true scores for these exams.

Table 9. Mean Indices of Internal Consistency and Standard Error of Measurement

Discipline	α	SEM at the Mean Score		SEM at the Cut Score	
		Raw	Scaled	Raw	Scaled
RAD	0.93	5.38	1.91	6.42	2.28
NMT	0.93	5.66	1.82	6.45	2.07
THR	0.90	5.44	2.03	6.26	2.34

Standard Error of Measurement

The standard error of measurement (SEM) is a type of standard deviation. SEM is the standard deviation of a hypothetical set of repeated measurements for a single individual. A common equation calculates the SEM using the reliability estimate, r_{XX} (α from Equation 1), and the standard deviation of exam scores, S_X, with the equation

$$\text{SEM} = S_X \sqrt{1 - r_{XX}} \ . \tag{2}$$

The above equation for SEM represents the mean SEM across all exam scores. SEM is not consistent, however, across the full range of scores, especially at the extremes. The SEM calculated at the cut score and the mean score will give a more accurate picture of the standard error. The equation for SEM at a particular score is

$$\text{SEM}_{\hat{X}} = \sqrt{\left(\frac{\hat{X}(k - \hat{X})}{k - 1} \right) \left(\frac{1 - r_{XX}}{1 - r_{21}} \right)}, \tag{3}$$

where $\hat{X}$ is a score value of interest,
k is the number of items (200 for each of these exams),
r_{XX} is the reliability of scores using Cronbach's α, and
r_{21} is the reliability of scores using Kuder-Richardson Equation 21 (Lord, 1955; Keats, 1957).

Table 9 provides the standard error of measurement for the mean score and the cut score in both raw and scaled score units using Equation 3.

Decision Consistency

ARRT administers criterion-referenced examinations as the basis of decisions to grant certification. Agreement indices quantify the consistency or reproducibility of those dichotomous (two option) decisions. Decision consistency in this case describes how consistently the examinations classify individuals into certified and not certified groups. When organizations base a pass/fail decision on a single exam score, there will be a small number of candidates who passed but should have failed (false positives) and a small number of candidates who failed but should have passed (false negatives). The threshold loss agreement indices used in this report focus on the consistency of classifications, treating all potential misclassification errors as equally serious.

The threshold loss indices assume a dichotomous, qualitative classification of candidates as certified or not certified based on a cut score. The methods were originally developed using two or more exam administrations for every examinee. Because multiple examinations are not practical, researchers developed alternative methods to estimate the indices with a single exam administration. This report uses a method developed by Subkoviak (1976) to estimate two threshold loss indices, p_0 and kappa. The estimation procedure assumes that a candidate's observed scores are independently and binomially distributed according to the number of exam items and the person's proportion-correct true score.

p_0 index

The p_0 index measures the overall consistency of pass/fail classifications. It is the proportion of individuals expected to be consistently classified as certified and not certified based on Subkoviak's (1976) method. The index is sensitive to the cut score, exam length, and score variability. For example, p_0 values will be smaller for cut scores near the mean of scores, because there are more people located near the mean than at the extremes if scores are normally distributed. The first column in Table 10 contains the p_0 values for each of the primary exam programs. Classification decisions based on ARRT's primary exams are consistent between 95% and 96% of the time. This is a high level of decision consistency.

Table 10. Threshold Loss Indices for 2012

Discipline	p_0	p_c	kappa
RAD	0.96	0.87	0.69
NMT	0.95	0.82	0.72
THR	0.95	0.85	0.67

Kappa

While high classification consistencies are good, it is possible that some or many of the correct classifications of certified or not certified were due to chance. For example, a person can correctly guess heads or tails at the flip of a coin a certain percentage of the time. These correct guesses are due purely to chance. Kappa is a statistical index that shows proportion of individuals consistently classified beyond that expected by chance. The equation for kappa is

$$k = \frac{p_0 - p_c}{1 - p_c}, \tag{4}$$

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.
The calculation for p_c is simply

$$p_c = (P_{Pass})^2 + (1 - P_{Pass})^2, \tag{5}$$

where P_{pass} is the proportion of people who pass the exam (Croker & Algina, 1986). Table 10 contains the kappa statistics for ARRT's exams. The kappa coefficient indicates that ARRT's exams consistently classify between 67% and 72% of the candidates above and beyond those already correctly classified by chance.

With regard to psychometric properties, ARRT's examinations are comparable to other well-developed certification examinations. ARRT's exam scores are reliable, with α coefficients above 0.90. The threshold loss indices indicate that most candidates are consistently classified as either certified or not certified. Maintaining a high quality examination program is a vital part of ARRT's mission of promoting high standards of patient care by recognizing qualified individuals in medical imaging, interventional procedures, and radiation therapy. The results from this technical report show that ARRT indeed continues to develop quality certification examinations.

References

Croker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.

Keats, J.A. (1957). Estimation of error variances of test scores. *Psychometrika, 2*, 29-41.

Lord, F.M. (1955). Estimating test reliability. *Educational and Psychological Measurement, 15*, 325-336.

Nunnally, J.C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Subkoviak, M.J. (1976). Estimating reliability from a single administration of a mastery test. *Journal of Educational Measurement, 13*, 265-276.