Annual Report of Examinations: Technical Appendix 2021

Introduction

This report summarizes the psychometric characteristics of ARRT's examination scores in Radiography (RAD), Nuclear Medicine Technology (NMT), Radiation Therapy (THR), Sonography (SON), and Magnetic Resonance Imaging (MRI) for the year 2021. This report is a companion document to the Annual Report of Examinations: Primary Eligibility Pathway.

The first section of this report contains information about the amount 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 and registration decisions that ARRT makes based on its examinations.

Updates

Starting in mid-2021, ARRT introduced a primary eligibility pathway for Vascular Sonography (VS). See the news article here for more information. Vascular Sonography will not appear in this year's report as this pathway was not utilized in 2021.

Information about Exam Durations

Most examination administrators, including ARRT, do not intend for exam administration time to be a major factor for candidates. Practical limitations, however, make it necessary to establish exam time limits. For RAD, NMT, THR, and MRI, candidates may take up to 210 minutes (3.5 hours) to answer 220 items (questions). For SON, candidates may take up to 390 minutes (6.5 hours) to answer 400 items. 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 if 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 candidates used to take the exams described in this report. Some sources (e.g., Nunnally, 1978) specify that an exam is unspeeded when at least 90% of candidates complete the exam within the allotted time. If results show that more than 10% of candidates require the full time, ARRT would consider re-evaluating existing time limits.

Table 1 contains a summary of the amount of time candidates spent on the exam. These and all other statistics reflect only first-time ARRT exam candidates. None of the statistics include state candidates or people retaking the exam after failing the initial attempt. This table indicates that THR candidates spent more time than their counterparts in RAD, NMT, and MRI. THR had the highest mean (average) time among the exams with 200 questions. SON took more time overall, but the time per item was lower than the other four disciplines.

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

Discipline	Number of Candidates	Minimum Time	Maximum Time	Mean Time	Standard Deviation
Radiography	12,255	46	210	146.36	40.00
Nuclear Medicine	414	50	210	144.70	41.61
Radiation Therapy	846	68	210	171.59	32.49
Sonography	486	79	390	232.30	70.03
Magnetic Resonance Imaging	2,401	42	210	140.79	41.85

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 94 minutes or less, and 20% completed it in 109 minutes or less. Continuing on the row, Table 2 shows that 90% of RAD candidates completed the exam in 201 minutes or less. Overall, most candidates completed their examinations within the established time limits. For all disciplines but Radiation Therapy, 90% or more of the candidates completed the exam in less than the allotted time. Radiation Therapy is being investigated internally regarding speededness. The other 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

	Cumulative Percentage of Candidates Completing the Exam								
Discipline	10%	20%	30%	40%	50%	60%	70%	80%	90%
Radiography	94	109	121	133	145	156	170	184	201
Nuclear Medicine	92	102	117	131	142	158	173	191	204
Radiation Therapy	123	141	156	166	176	188	198	205	210
Sonography	144	167	187	204	228	247	273	295	334
Magnetic Resonance Imaging	85	100	114	125	140	154	168	185	200

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. The total score consists of 200 items for RAD, NMT, THR, and MRI. The total score consists of 360 items for SON. There are also additional unscored "pilot" items on each exam.

Table 3. Descriptive Statistics of Raw Scores

Discipline	**Minimum**	**Maximum**	**Mean**	**Standard Deviation**
Radiography	43	198	154.98	21.51
Nuclear Medicine	71	191	143.92	24.16
Radiation Therapy	79	195	152.35	20.08
Sonography	127	344	254.12	44.72
Magnetic Resonance Imaging	49	198	147.03	24.38

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. 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**
Radiography	40	99	82.27	8.24
Nuclear Medicine	57	97	80.98	8.06
Radiation Therapy	52	98	80.67	7.92
Sonography	51	96	77.43	9.28
Magnetic Resonance Imaging	43	99	79.84	9.06

In order to convert raw scores to scaled scores, ARRT determines the difficulty of an exam form. Each exam consists of items that were used on previous exams. ARRT uses the Rasch model to track the difficulty levels of individual exam items and, consequently, whole exam forms. Each item has a Rasch difficulty statistic indicating the probability of a candidate answering correctly.

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).

$$a = (99.49 - 74.50) / (200 - 130) = 0.357$$

$$b = 74.50 - (a \times 130) = 74.50 - (0.357 \times 130) = 28.09$$

For this hypothetical form, the scaling coefficients would be $a = 0.357$ 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

$$\text{scaled score} = (\text{raw score} \times 0.357) + 28.09 = (131 \times 0.357) + 28.09 = 74.857,$$

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 to a scaled score of 74, which is a failing score.

Table 5 contains the candidate passing percentages for exams taken by primary pathway candidates. This information is also in the Annual Report of Examinations: Primary Eligibility Pathway report but is repeated here because of its importance.

Table 5. Pass Percentages for First-Time Candidates

Discipline	***Pass Percentage***
Radiography	83.76
Nuclear Medicine	79.47
Radiation Therapy	79.79
Sonography	55.56
Magnetic Resonance Imaging	74.59

Descriptive Statistics for Section Scores

In addition to the total scaled score, ARRT reports individual section scores that correspond to content areas as outlined in the content specifications of each exam. The primary purpose of the section scores is to provide general information to candidates regarding their strengths and weaknesses in particular content categories. For SON only, candidates must pass both the Abdomen and OB/GYN sections with a section scaled score of 7.5 in addition to passing the total test (scaled score of 75). ARRT reports section scores on a scale from 0.1 to 9.9 in one-tenth point intervals.

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 10 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. Reliability is discussed in more detail later in this report.

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 10, 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 account for the unreliability of the section scores and give a sense of the magnitude of the correlations under the condition of perfect reliability. The high correlations after correction among many of the section scaled scores indicate a high degree of common variance among these scores.

Table 6. RAD Section Score Correlation Matrix and Statistics

Content Area	***PC1***	***S1***	***S2***	***IP1***	***IP2***	***P1***	***P2***	***P3***
PC1		0.49	0.57	0.56	0.57	0.50	0.53	0.49
S1	*0.76*		0.65	0.58	0.63	0.53	0.54	0.55
S2	*0.82*	*0.94*		0.67	0.67	0.60	0.60	0.59
IP1	*0.83*	*0.87*	*0.92*		0.67	0.56	0.59	0.57
IP2	*0.87*	*0.96*	*0.96*	*0.98*		0.54	0.57	0.58
P1	*0.78*	*0.83*	*0.87*	*0.85*	*0.83*		0.61	0.61
P2	*0.84*	*0.86*	*0.88*	*0.90*	*0.90*	*0.97*		0.61
P3	*0.73*	*0.82*	*0.82*	*0.82*	*0.85*	*0.93*	*0.93*	

Note: Italicized correlations below the diagonal are corrected for unreliability

STATISTIC	***PC1***	***S1***	***S2***	***IP1***	***IP2***	***P1***	***P2***	***P3***
No. Items	33	21	29	26	25	18	20	28
Mean ss	8.45	8.23	8.11	8.15	8.09	8.33	8.16	8.31
Sd ss	0.82	1.10	1.07	1.14	1.04	1.13	1.08	0.99
Mean raw	26.48	17.01	23.52	16.05	21.94	14.19	16.07	19.71
SD raw	3.54	3.15	4.31	3.14	3.92	2.69	2.95	3.28
Reliability	0.64	0.64	0.75	0.70	0.66	0.63	0.61	0.69

RAD Section Key:

Abbreviation	***Section Name***
PC	Patient Care
S	Safety
IP	Image Production
P	Procedures
PC1	Patient Interactions and Management
S1	Radiation Physics and Radiobiology
S2	Radiation Protection
IP1	Image Acquisition and Technical Evaluation
IP2	Equipment Operation and Quality Assurance
P1	Head, Spine, and Pelvis Procedures
P2	Thorax and Abdomen Procedures
P3	Extremity Procedures

Table 7. NMT Section Score Correlation Matrix and Statistics

Content Area	PC1	S1	IP1	P1	P2	P3	P4	P5
PC1		0.37	0.46	0.36	0.41	0.41	0.48	0.50
S1	*0.60*		0.70	0.64	0.63	0.65	0.59	0.64
IP1	*0.71*	*0.99*		0.66	0.67	0.68	0.66	0.71
P1	*0.57*	*0.94*	*0.93*		0.66	0.69	0.67	0.62
P2	*0.63*	*0.9*	*0.91*	*0.93*		0.72	0.67	0.73
P3	*0.66*	*0.97*	*0.97*	*1.01*	*1.02*		0.71	0.74
P4	*0.75*	*0.85*	*0.91*	*0.95*	*0.93*	*1.02*		0.69
P5	*0.78*	*0.92*	*0.98*	*0.89*	*1.01*	*1.07*	*0.95*	

Note: Italicized correlations below the diagonal are corrected for unreliability

Statistic	PC1	S1	IP1	P1	P2	P3	P4	P5
No. Items	24	25	33	28	25	25	18	22
Mean SS	8.18	7.97	8.09	7.92	8.27	7.69	8.32	8.36
SD SS	0.91	0.97	0.90	0.94	0.94	1.07	1.08	1.01
Mean Raw	14.64	15.57	27.33	16.65	17.93	18.46	15.08	18.25
SD Raw	2.72	3.22	5.15	3.43	3.41	4.50	3.28	3.65
Reliability	0.57	0.68	0.73	0.68	0.73	0.68	0.72	0.72

RAD Section Key:

Abbreviation	Section Name
PC	Patient Care
S	Safety
IP	Image Production
P	Procedures
PC1	Patient Interactions and Management
S1	Radiation Physics, Radiobiology, and Regulations
IP1	Instrumentation
P1	Radionuclides and Radiopharmaceuticals
P2	Cardiac Procedures
P3	Endocrine and Oncology Procedures
P4	Gastrointestinal and Genitourinary Procedures
P5	Other Imaging Procedures

Table 8. THR Section Score Correlation Matrix and Statistics

Content Area	***PC1***	***PC2***	***S1***	***S2***	***P1***	***P2***	***P3***	***P4***
PC1		0.48	0.43	0.48	0.49	0.48	0.53	0.53
PC2	*0.87*		0.52	0.53	0.59	0.50	0.54	0.57
S1	*0.73*	*0.96*		0.63	0.58	0.51	0.60	0.61
S2	*0.79*	*0.94*	*1.04*		0.60	0.53	0.62	0.61
P1	*0.80*	*1.04*	*0.96*	*0.95*		0.53	0.52	0.56
P2	*0.86*	*0.97*	*0.93*	*0.92*	*0.92*		0.50	0.58
P3	*0.82*	*0.90*	*0.94*	*0.92*	*0.77*	*0.82*		0.69
P4	*0.82*	*0.95*	*0.96*	*0.91*	*0.84*	*0.94*	*0.97*	

Note: Italicized correlations below the diagonal are corrected for unreliability

Statistic	***PC1***	***PC2***	***S1***	***S2***	***P1***	***P2***	***P3***	***P4***
No. Items	29	17	21	30	26	18	24	35
Mean SS	8.57	8.11	7.68	7.66	8.07	8.24	7.76	8.35
SD SS	0.84	0.98	1.17	1.02	0.99	1.05	1.22	0.90
Mean Raw	20.64	16.87	14.2	20.62	19.89	14.1	17.34	28.7
SD Raw	2.69	2.74	2.94	3.74	3.3	2.4	3.71	4.09
Reliability	0.59	0.51	0.58	0.63	0.63	0.53	0.71	0.71

THR Section Key:

Abbreviation	***Section Name***
PC	Patient Care
S	Safety
P	Procedures
PC1	Patient Interactions
PC2	Patient and Medical Record Management
S1	Radiation Physics, Equipment, and Quality Assurance
S2	Radiation Protection
P1	Treatment Sites and Tumors
P2	Treatment Volume Localization
P3	Prescription and Dose Calculation
P4	Treatments

Table 9. SON Section Score Correlation Matrix and Statistics

Content Area	**PC1**	**IP1**	**IP2**	**IP3**	**P1**	**P2**	**P3**
PC1		0.46	0.47	0.49	0.56	0.42	0.50
IP1	*0.46*		0.82	0.70	0.63	0.50	0.56
IP2	*0.47*	*0.82*		0.69	0.63	0.49	0.57
IP3	*0.49*	*0.70*	*0.69*		0.67	0.51	0.59
P1	*0.56*	*0.63*	*0.63*	*0.67*		0.65	0.79
P2	*0.42*	*0.50*	*0.49*	*0.51*	*0.65*		0.72
P3	*0.50*	*0.56*	*0.57*	*0.59*	*0.79*	*0.72*	

Note: Italicized correlations below the diagonal are corrected for unreliability

Statistic	**PC1**	**IP1**	**IP2**	**IP3**	**P1**	**P2**	**P3**
No. Items	29	50	44	21	75	109	32
Mean SS	8.04	7.37	7.55	8.28	7.76	7.83	7.64
SD SS	0.95	1.10	1.08	1.17	1.02	1.11	1.13
Mean Raw	21.62	32.84	29.91	16.33	53.20	78.22	22.16
SD Raw	3.67	7.47	6.33	3.29	10.63	16.17	4.85
Reliability	0.66	0.84	0.81	0.72	0.89	0.93	0.75

SON Section Key:

Abbreviation	**Section Name**
PC	Patient Care
IP	Image Production
P	Procedures
PC1	Patient Interactions and Management
IP1	Basic Principles of Ultrasound
IP2	Image Formation
IP3	Evaluation and Selection of Representative Images
P1	Abdomen
P2	OB/GYN
P3	Superficial Structures and Other Sonographic Procedures

Table 10. MRI Section Score Correlation Matrix and Statistics

Content Area	***PC1***	***S1***	***IP1***	***IP2***	***IP3***	***P1***	***P2***	***P3***
PC1		0.53	0.50	0.45	0.48	0.50	0.46	0.45
S1	*0.89*		0.62	0.60	0.60	0.55	0.47	0.51
IP1	*0.74*	*0.89*		0.77	0.77	0.63	0.52	0.57
IP2	*0.65*	*0.86*	*0.96*		0.77	0.64	0.53	0.60
IP3	*0.71*	*0.87*	*0.98*	*0.98*		0.62	0.52	0.56
P1	*0.78*	*0.85*	*0.84*	*0.85*	*0.85*		0.60	0.61
P2	*0.76*	*0.77*	*0.74*	*0.76*	*0.75*	*0.92*		0.55
P3	*0.75*	*0.84*	*0.82*	*0.86*	*0.81*	*0.93*	*0.89*	

Note: Italicized correlations below the diagonal are corrected for unreliability

Statistic	***PC1***	***S1***	***IP1***	***IP2***	***IP3***	***P1***	***P2***	***P3***
No. Items	18	20	39	36	30	25	15	17
Mean SS	8.12	8.06	7.91	8.02	7.81	8.23	7.72	8.08
SD SS	1.06	1.05	1.07	1.11	1.18	1.06	1.30	1.15
Mean Raw	13.58	14.91	28.28	26.63	21.34	19.11	10.44	12.74
SD Raw	2.59	2.83	5.64	5.43	4.73	3.55	2.62	2.66
Reliability	0.58	0.60	0.80	0.81	0.78	0.70	0.62	0.62

MRI Section Key:

Abbreviation	***Section Name***
PC	Patient Care
S	Safety
IP	Image Production
P	Procedures
PC1	Patient Interactions and Management
S1	MRI Screening and Safety
IP1	Physical Principles of Image Formation
IP2	Sequence Parameters and Options
IP3	Data Acquisition, Processing, and Storage
P1	Neurological
P2	Body
P3	Musculoskeletal

When interpreting the correlations in Tables 6 through 10, 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 the candidate's 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 may not have sufficient reliability to make those decisions. A notable exception to this is SON. ARRT does base pass/fail decisions on the Abdomen and OB/GYN sections of that exam, and the reliability of those section scores is quite high.

Reliability of Exam Scores

Reliability refers to the repeatability and consistency of exam scores. A candidate 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 candidate 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 candidate's knowledge or if the exam scores are unreliable.

Reliability also describes how well candidates' observed scores on an exam approximate their "true" scores. A candidate's true score may be defined as the mean of their observed scores from a very large number of examinations. The 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 candidate's score, the standard error of measurement (SEM) describes the expected variation of each candidate's observed score around that candidate's true score.

Coefficient Alpha

The equation for Cronbach's coefficient *α* is

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

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 11 contains the reliability estimates for RAD, NMT, THR, SON, and MRI. Recalling that reliability coefficients range from 0.0 to 1.0, one can see that the reliability estimates for the exam scores are quite high at 0.87 or greater. These high reliability estimates mean that observed scores for these exams likely correspond quite closely to true scores for these exams.

Table 11. Mean Indices of Internal Consistency and Standard Error of Measurement

		SEM at the Mean Score		***SEM at the Cut Score***	
Discipline	**α**	**Raw**	**Scaled**	**Raw**	**Scaled**
Radiography	0.94	5.61	2.15	6.31	2.41
Nuclear Medicine	0.94	5.98	1.98	6.45	2.14
Radiation Therapy	0.92	5.66	2.23	6.18	2.44
Sonography	0.87	6.42	1.34	6.64	1.38
Magnetic Resonance Imaging	0.95	5.92	2.20	6.34	2.35

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,
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 11 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 examinations with criterion-referenced cut score standards as the basis of decisions to grant certification and registration. 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 registered and not certified and registered 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 and registered or not certified and registered based on a cut score. The methods were originally developed using two or more exam administrations for every candidate. 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, p0 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 candidate'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 registered and not certified and registered 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 12 contains the p_0 values for each of the exams that this report covers. Classification decisions based on these exams are consistent between 90% and 93% of the time. This is a high level of decision consistency.

Table 12. Threshold Loss Indices

Discipline	p_0	p_c	***kappa***
Radiography	0.93	0.73	0.74
Nuclear Medicine	0.91	0.57	0.79
Radiation Therapy	0.92	0.68	0.75
*Sonography**	0.90	0.54	0.78
Magnetic Resonance Imaging	0.92	0.62	0.79

* The p_0 statistic for SON makes a statistical adjustment to Subkoviak's (1976) method that accounts for the necessity to pass the overall exam, the Abdomen section, and the OB/GYN section.

Kappa

While high classification consistencies are good, it is possible that some or many of the correct classifications of certified and registered or not certified and registered 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 the 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 and registered/not certified and registered 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 74% and 79% 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 examinations. ARRT's exam scores are reliable, with *α* coefficients at or above .87. The threshold loss indices indicate that most candidates are consistently classified as either certified and registered or not certified and registered. 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 examinations.

References

Crocker, 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.