INTRODUCTION

This report summarizes the results of the 2019 Registered Radiologist Assistant (R.R.A.®) examinations developed and administered by the ARRT. The purpose of the R.R.A. examination is to assess the knowledge and cognitive skills underlying the performance of the clinical activities and imaging procedures required of the radiologist assistant. The examinations administered were assembled according to the job analysis, ELCA, and content specifications that were effective beginning in July 2018.

The exams consist of a selected response session and a constructed response session given over a 6-hour period. The selected response component is primarily composed of traditional multiple-choice questions. Candidates identify the correct answer from the choices listed. The goal of the selected response component is to evaluate breadth of knowledge and cognitive skills. The constructed response component consists of patient cases followed by essay questions. Candidates then construct their answer instead of choosing from a list. The constructed response questions evaluate depth of knowledge and clinical reasoning skills. These two assessment formats are further described below.

Selected Response. The selected response (SR) component consists of 200 scored plus 20 pilot questions. Candidates are allowed up to 3.5 hours to complete the 220 questions. The following tables list the major topics covered. Table 1 details the topics covered by the 2019 exams. Detailed content specifications are available at arrt.org.

Table 1. Content Categories for Selected Response Sections	**No. of Items**
Patient Care	60
Patient Management (34)	
Pharmacology (26)	
Safety	25
Patient Safety, Radiation Protection and Equipment Operation (25)	
Procedures	115
Abdominal Section (43)	
Thoracic Section (29)	
Musculoskeletal and Endocrine Sections (20)	
Neurological, Vascular and Lymphatic Sections (23)	
Total	200

Most selected response items are standard multiple-choice questions that require just one best answer. However, this format also includes *select multiple*, *sorted list,* and *hot area* items*.* A select multiple is a question followed by a list of four to eight response options, and candidates select all that are correct. The sorted list format presents a list of four to ten options, and candidates are required to place the options in correct sequence by using the mouse to "drag-and-drop" them so that they are in proper order. A hot area is a question accompanied by a medical image or other graphic that requires candidates to use the mouse to identify a location or region on the exhibit.

Constructed Response. This component consists of two case studies, each followed by four to seven essay questions. Candidates are given up to two and a half hours to complete the cases. Each case opens with a brief scenario describing a patient in need of radiology-related services. The scenario may indicate the presenting complaint, patient history, results of diagnostic tests, suspected diagnosis, and previous treatments. The questions can address a variety of topics related to the case, such as explaining how to perform a procedure; discussing contraindications; reviewing images and suggesting preliminary observations; identifying whether additional diagnostic studies might be necessary; describing anatomy; or explaining follow-up care to a particular patient. In addition to the essay questions, most cases also include a few selected response questions.

Cases included on the R.R.A. exam were sampled from the domain of the 13 mandatory procedures that students are expected to have successfully completed during the course of their clinical preceptorship based on the July 2018 job analysis, ELCA, and content specifications. The 13 procedures are listed in Table 2. A sample case study is available at arrt.org.

Table 2. Content for Case-Based Essays

Abdominal Procedures
- *General Abdomen*
  1. Paracentesis
- *Gastrointestinal*
  2. Esophageal study
  3. Swallowing function study
  4. Upper GI study
  5. Small bowel study
  6. Enema with barium, air-, or water-soluble contrast
  7. Nasogastric/enteric and orogastric/enteric tube placement
- *Urinary*
  8. Cystography, voiding cystography or voiding cystourethrography

Thoracic procedures
- *Pulmonary*
  9. Thoracentesis

Musculoskeletal and Endocrine Procedures
- *Musculoskeletal*
  10. Arthrogram (shoulder or hip)

Neurological, Vascular and Lymphatic Procedures
- *Neurological*
  11. Lumbar puncture with or without contrast
  12. Cervical, thoracic, or lumbar myelography
- *Vascular and Lymphatic*
  13. Peripherally inserted central catheter (PICC) placement

INTERPRETING SCORES

Total Scaled Score. The ARRT uses scaled scores to report exam results. A total scaled score can range from 1 to 99, and a total scaled score of 75 is required to pass an examination.

Scaled scores are desirable because they take into account the difficulty of a particular exam compared to earlier versions. Raw scores (i.e., number correct, or percent correct) have limited use because they cannot be compared from one version of an exam to the next. This lack of comparability exists because one version of an exam might be slightly easier or slightly more difficult than a previous version. Scaled scores take into account any differences in difficulty between two or more versions of an exam. A scaled score of 75 represents the same level of exam performance, regardless of which version an examinee takes.

Scaled scores are sometimes mistaken for percent correct scores. This confusion probably arises because both scaled scores and percentages have a similar range of values. A scaled score of 75 does not mean that someone correctly answered 75% of the test questions.

Section and Essay Scores. Performance on each section of the exam is also reported using scaled scores. Pass-fail decisions are not based on section scores; the scaling of section scores is intended to help candidates evaluate their performance on different parts of the test.

Section scores can range from 0.1 to 9.9 and are reported in one-tenth point intervals (e.g., 8.1, 8.6). Section scores are intentionally placed on a narrower scale because they are often based on a small number of test questions. Therefore, section scores are not as reliable as the total scaled score and should be interpreted with some caution.

Passing Score. A scaled score of 75 is required to pass all ARRT exams. This pass-fail point, called the "cut score", is reviewed periodically by an advisory committee and established by ARRT's Board of Trustees through a process called standard setting. During the standard setting process, the advisory committee consisting of R.R.A.s, educators, and radiologists conducts structured activities to arrive at a recommended cut score. The Board reviews the results of these activities to establish a final cut score. The cut score represents the standard of performance required to obtain certification. Those who meet or exceed the standard pass the exam.

One may ask how many questions need to be answered correctly to achieve a scaled score of 75. The answer depends on the difficulty of the particular form that was taken. For most R.R.A. test forms, a scaled score of 75 corresponds to about 65% to 70% correct for the selected response portion, and about 60% to 68% correct for the case-based essay component. Again, it is important to note that the cut score will vary slightly based on the difficulty of a particular test form. For example, if a July test form is 2 points more difficult than the previous January test form, then the *raw* passing score for the July exam would be two-points lower. However, the *scaled* passing score would remain at 75. Test form difficulty is monitored through a process known as statistical equating.

Scoring Case-Based Essays. Each essay question is worth 3 to 6 points depending on the complexity of the question, as well as the length and detail expected in the response. For example, a question that asks a candidate to *list* four types of imaging studies that could be used to evaluate a suspected pathology might be worth 3 points, while a question that asks the candidate to list four types of imaging studies *and explain* the diagnostic utility of each would be worth 6 points. Candidates are informed of the point value of each question during exam administration. Regardless of the exact number of points on the essay, the total score is always computed such that the essays account for 25% of the total, while the selected response sections account for 75% of the total.

Essays are graded by the Essay Evaluation Committee (EEC). The ARRT recognizes that essay scoring has an element of subjectivity. This potential limitation is addressed through: (a) the use of detailed scoring rubrics[1] specific to each case and question; (b) orienting evaluators by holding a practice scoring session prior to actual grading; (c) having each essay graded by three evaluators; (d) maintaining anonymity; (e) arranging essay responses such that evaluators grade all responses to a particular question before scoring the responses to the next question; (f) randomizing responses so that no single candidate is always graded first, last, or in the middle; (g) discussing essays for which initial scores exhibit disagreement; (h) providing evaluators feedback regarding their ratings.

Having each essay graded by three individuals, in combination with the scoring process outlined above, helps to ensure the reliability and validity of essay scores.[2]

[1]A scoring rubric is a detailed set of scoring guidelines. The rubric for each question occupies about a full page. The first part of the rubric delineates the information that the response to each question should or could contain. The second part presents rules for assigning scores to the information present in a response.

[2] Additional analyses indicate that essay evaluators (graders) exhibit high levels of agreement for the scores they assign as an outcome from this comprehensive scoring process. The median interrater correlations for the two exam administrations were .87 and .67.

EXAMINATION RESULTS

Overall Performance Statistics. A total of 16 R.R.A examinations were administered. Of these, 13 were first-time candidates. 11 of the 13 first-time candidates passed, for a first-time pass rate of 84.6% for the year.

School Performance. The 13 first-time candidates represented five R.R.A. educational programs. The number of graduates from each program varied with the programs having between one and six total graduates. For these programs, the mean scaled scores ranged from 76.4 to 87. The number of candidates and programs at this point in time is too small for these data to be described in further detail.

Score Trends. The ARRT routinely monitors exam statistics over time for all of its certification programs. Recent available data is summarized in Table 4. The results are encouraging because they indicate that mean scores have been relatively stable over the course of the past several years when taking into account both administrations together.

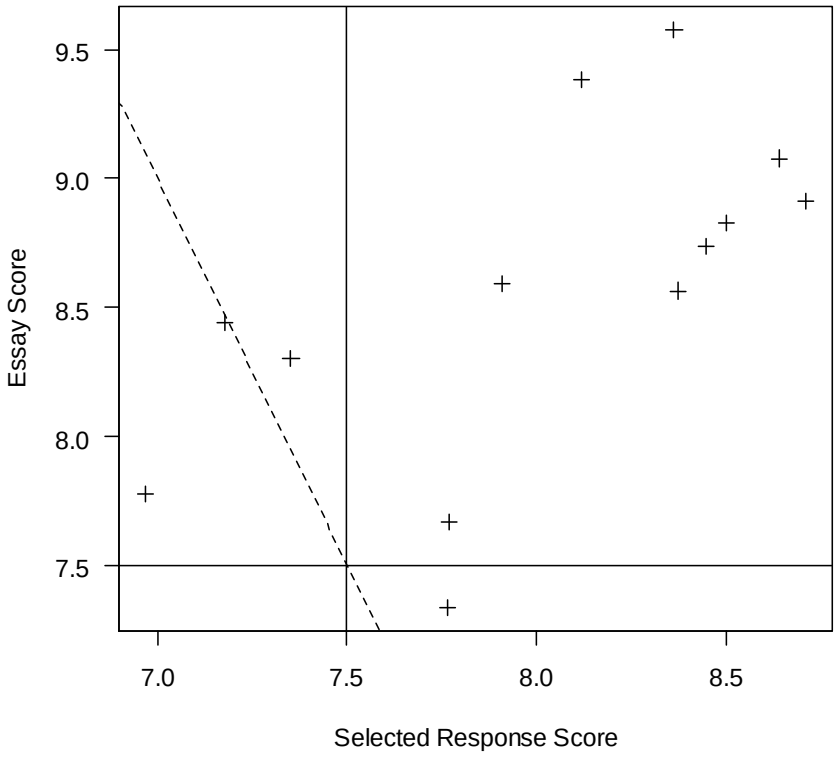
Table 3. Examination Statistics
First-Time Candidates (N=13)

	Scaled Score			
Section	**Mean**	**SD**	**Min**	**Max**
Patient Mgmt	8.2	0.6	7.1	9.2
Pharmacology	8.0	0.9	6.3	9.3
PS, RP and EO	7.3	0.7	6.2	8.7
Abdominal	8.4	0.7	7.4	9.4
Thoracic	8.0	0.8	6.4	9.1
MSK/Endocrine	7.9	0.7	6.3	8.8
N/V/L	7.4	1.0	5.6	8.6
Case Essay	8.5	0.7	7.3	9.5
Total Score	**80.8**	**5.6**	**71**	**87**

Table 4.
2005 - 2019 Summary Statistics
for First-Time, Repeat, and All Candidates

Year	Group	N	Mean	% Pass
2013	First	32	79.3	78.1%
	Repeat	18	72.6	38.9%
	Total	50	76.9	64.0%
2014	First	33	77.2	63.6%
	Repeat	20	72.7	30.0%
	Total	53	75.5	50.9%
2015	First	28	79.9	78.6%
	Repeat	10	71.6	30.0%
	Total	38	77.7	65.8%
2016	First	16	79.9	75.0%
	Repeat	6	75.8	66.7%
	Total	22	78.8	72.7%
2017	First	20	79.9	85.0%
	Repeat	1	NA	NA
	Total	21	NA	NA
2018	First	15	78.3	73.3%
	Repeat	5	73.4	40.0%
	Total	20	77.1	65.0%
2019	First	13	80.8	84.6%
	Repeat	3	74.7	66.7%
	Total	16	79.7	81.3%
TOTAL	First	481	79.4	78.2%
2005-2019	Repeat	132	73.1	37.1%
	Total	613	78.0	69.3%

Figure 1. Scatterplot of Essay Scores with Selected Response Scores.



Case-Based Essays. The ARRT closely monitors the essay component of the exam. Although scores on case-based essay questions are combined with scores on the selected response component to obtain final scores, it is still useful to inspect scores on both sections to determine if they are related. The scatterplot in Figure 1 displays the joint distribution of scores on the two parts of the exam.

Each point in Figure 1 represents a candidate's score on both parts of the exam. The diagonal dotted line represents the approximate overall scaled score passing value of 75 resulting from differing combinations of scores on the two parts of the exam. The general trend is that candidates who do well on one component tend to score well on the other, and vice versa. The correlation (*r*) between the two sets of scores is 0.63. If the scores were perfectly correlated (fell on a straight line, $r = 1.0$), then the essays would be contributing no additional information beyond what is provided by the selected response component. In contrast, one would have to question the validity of the essay section if the two sets of scores were completely unrelated (appeared more like a circle, $r = 0.0$).

The pattern depicted in the graph is reassuring. It indicates that while the two components have some things in common, each still contributes unique information to the measurement of candidate proficiency.

Correlation coefficients are especially susceptible to the type of statistical error associated with small sample sizes. Therefore, the data in this chart, and the corresponding *r* value, should be interpreted with considerable caution

SCORE RELIABILITY

Reliability refers to the consistency or dependability of the measurements or scores obtained for some entity. Just as physical measurements (e.g., blood pressure readings) can be different for an individual on two occasions, so can a candidate's score on a certification exam. Reliability describes the extent to which a candidate's score approximates his or her true score. Coefficient alpha is the most common way to quantify a test's reliability. The converse of reliability is measurement error.

The standard error of measurement (SEM) indicates the amount that a person's score is expected to differ

on repeated measurements. Both coefficient alpha and the SEM are reported here.

Test scores on certification exams are ultimately used to make certification decisions. Therefore, it makes sense to evaluate test scores in terms of the dependability of those decisions. This report also documents the level of consistency for the pass-fail decisions made on the basis of scores on the R.R.A. exam.

Calculating reliability indices for a multiple-choice test is fairly straightforward, while calculating them for an essay exam is considerably more involved. Determining the reliability for scores obtained by combining scores on the two types of exams adds additional complexity. All of these calculations require sample sizes in the 100s in order to get good, stable estimates of reliability. Because reliability coefficients computed on the R.R.A. exam are based on substantially smaller samples, the coefficients may not be very stable. However, it is still useful to compute these values in an effort to detect any significant problems with the test scores.

Reliability and SEM. Table 5 indicates the reliability coefficients associated with each part of the exam. These values are based on first-time examinees on the January exam (N=4) and July exam (N=9). Cronbach's alpha for each entire exam was calculated factoring in the added weighting for the essay items (25% of the exam points). As indicated in Table 5, the level of reliability for the combined score – the score upon which pass-fail decisions are based – is high. The reliability estimate is somewhat lower in January due to very low variation in scores (see Table 3).

It is also instructive to evaluate the reliability coefficients for the individual components (i.e., selected response, essay). The reliabilities for just the selected response component are in the acceptable range for a test that is used to make pass-fail decisions. That is, if the R.R.A. exam did not include a case-based essay component, the scores would still possess adequate reliability.

Table 5. Reliability Coefficients

Exam Component	January	July
Selected Response	.63	.92
Essay	.65	.73
Combined	.77	.92

The reliabilities for the case-based essay component are near to or lower than the selected response component, as expected. The primary reason is that the case-based essay component consists of far fewer questions, and reliability is determined mostly by the number of questions on an exam. The level of reliability for the case-based essay component suggests that pass-fail decisions not be made on the basis of the case-based essay scores alone. This is a primary reason that ARRT combines the case-based essay and selected response scores into a single total score.

The conventional standard error of measurement (SEM) is estimated by the formula:

$$SEM = \sigma_c \sqrt{1 - r_{cc}} \ ,$$

where σ_c is the standard deviation of the combined scores. For the January and July test administrations, the standard deviations of the scaled scores were 4.1 and 6.3 (recall that scaled scores range from 1 to 99, with 75 defined as passing). The corresponding SEMs for the two R.R.A. exam administration dates are 1.97 and 1.78.

Decision Consistency. Decision consistency quantifies the agreement of classification decisions (certified vs. not certified) based on repeated test administrations. Since it is not practical to have all candidates take the test on multiple occasions, methods have been developed to estimate decision consistency using data from a single test administration. In this report, a method developed by Subkoviak (1976) was used to estimate two threshold loss indices, p-naught (p_0) and kappa (k).

The p_0 index measures the overall consistency of pass-fail classifications. It is the proportion of individuals consistently classified (certified or not) on repeated measurements. The index is sensitive to the cut-off score, test length, and score variability. Kappa is the proportion of individuals consistently classified beyond that expected by chance. It is calculated by:

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

where p_0 is the overall consistency of classifications, and p_c is the proportion of consistent classifications that would be expected by chance. Note that p_c is undefined if all candidates pass the exam.

Table 7. Consistency Indices

Consistency Index	January	July
p_0	.92	.85
p_c	NA	.65
k	NA	.58

Table 7 indicates that candidates would be consistently classified as certified or not certified 85%, and 92% of the time in a theoretical two test situation. Given the very small numbers of examinees taking the R.R.A. examination and the mixed item format these values would seem to be reasonable. All summary statistics for the R.R.A. examination, however, should be interpreted cautiously given the small volume of candidates taking the exams.

Concluding Comments

The numbers of candidates for this year's report continue to argue for a very cautious interpretation of the data presented herein. As the R.R.A. certification program continues to evolve, it is hoped that future reports will be based upon exam administration sample sizes that are larger as would be required to complete and document more extensive analyses of the exam data.