The Scientific Reasoning Test, Version 9 (SR-9)

Test Manual

Madison Assessment

The Center For Assessment & Research Studies

Donna L. Sundre

2008

MSC 6806
Harrisonburg, VA 22807
540.568.6706 Phone
540.568.7878 Fax
assessment@jmu.edu
www.jmu.edu/assessment

Table of Contents

The Scientific Reasoning Test, Version 9 (SR-9)

The SR-9 Test Manual

Section 1. The Nature of the Instrument

The Scientific Reasoning Test, Version 9 (SR-9) is a 49-item multiple-choice test developed by science and mathematics university faculty. This instrument was designed to assess the scientific reasoning skills that college students may obtain through a general education curriculum. The Scientific Reasoning Test is currently in its ninth edition and reflects development spanning over the last decade.

Section 2. Intended Use

2.1. Appropriate and inappropriate uses and interpretations

This instrument was designed to assist departments in making improvements to curriculum, programs, and future iterations of their assessment processes. It was created to demonstrate student learning as a result of participation in the scientific components of general education programs. The SR-9 was developed for use at the programmatic level. Thus, any inferences made about learning or mastery should be made *only* in the aggregate.

The SR-9 was *not* designed for making decisions about individual students. Currently, psychometric properties are not sufficient to support high-stakes classifications for individuals (please refer to section 5.2 -- Evidence of reliability). This instrument was also not intended as a vehicle for providing individual students with feedback about their mastery of scientific reasoning. Institutions may choose to provide their students with individual feedback, but results should *not* be used to make any type of high-stakes classification decisions. According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2000), test users are responsible for collecting validity evidence for any uses of the test other than those recommended here.

The data collected with the SR-9 can be used to provide information about student learning that can inform improvements to a general education science curriculum. The results of the SR-9 can also be utilized to meet the State Council of Higher Education in Virginia's (SCHEV) requirements. SCHEV has mandated that all funded institutions must report on student competencies in the areas of scientific and quantitative reasoning. The SR-9 can be used to report on scientific reasoning. If an institution needs to assess quantitative reasoning, the Quantitative Reasoning Test, Version 9 (QR-9) or the Natural World Test, Version 9 (NW-9, which assesses both scientific and quantitative reasoning) may be used.

2.2. Target population

The SR-9 is intended to measure learning in scientific reasoning for undergraduate college students. Since items were designed to be content-free, this instrument should be appropriate for students in any general education science curriculum. Although the SR-9 was designed with all undergraduate college students in mind, it is important to take note of the normative sample when using this instrument. Please refer to Section 5.4 (Norming) for more information.

2.3. Qualifications of users

Test users must be trained to administer assessments in standardized conditions. The Proctor qualifications and training section of this manual (Section 4.1) provides more information about how proctors can be trained for test administration. JMU will complete scoring[1] for this instrument. In addition, test users should be knowledgeable about how to interpret the statistical results from the test and how to make appropriate inferences about the program using the results. Test users who do not have a measurement background or do not have in depth knowledge of the program are encouraged to consult with colleagues who have the necessary knowledge.

[1] Scoring (correct/incorrect) includes all SR-9 items and, if desired, up to 20 additional items added to the instrument by those institutions seeking to further evaluate the construct.

Section 3. Test Development

3.1. Academic and theoretical basis

The SR-9 was designed to evaluate student learning in six general education objectives related to scientific reasoning. The first has to do with distinguishing science from pseudo-science, and methods of inquiry which allow individuals to arrive at sound conclusions. The second student-learning objective assesses whether students can use theories and models to understand and make predictions. The third and fourth objectives regard the interdependence of applied research, basic research, and technology and their impact on society. Students should also be able to articulate how scientific developments affect social and ethical issues. Next, students should have knowledge of experimental and research design to test hypotheses. The final objective regards evaluating the credibility, use and misuse of scientific and mathematical information.

3.2. Item type selection

All SR-9 items use a selected-response format, with the number of response options ranging from two to four. The items were written as such to ease scoring, to maintain objective scoring, and to minimize test-taker fatigue. Most items follow a typical multiple choice format, in which an item stem is followed by alternative responses consisting of the correct answer and at least one distracter. An effort was made for items on the SR-9 to have only three alternative responses, with two high quality distracters. Items with more than three alternative responses were often part of a set of items in which the same alternatives were provided for each (similar to a matching item design).

3.3. Item pool and scale development process

Open-ended interviews with mathematics and science faculty members provided information about the construct and objectives. Guided by this information, items were written by faculty in direct relation to the objectives. A back translation (Dawis, 1987) was subsequently conducted to determine how items matched back to the appropriate objective.

Starting with the eighth version of the Scientific Reasoning test, faculty also completed a content alignment activity to establish that the test items did correspond with the objectives which they were intended to assess. In contrast to the back translation, the content alignment exercises asks each panelist to consider each objective, one at a time, then to go through the exam to locate items which might contribute to the measurement of the objective. The content alignment method is preferable to the back translation since judges tend to assign an item to a single objective during the back translation procedure, even if the item is a good match to more than one objective (Miller, Setzer, Sundre & Zeng, 2006).

Once the back translation and content alignment activities were completed, the items were administered to a small pilot study group to determine how easy the instrument's instructions were to follow and to examine test format, length of test, and the appropriateness of the items to the college student population. Next, the test was administered to a random sample of first-year college students and then to a randomly selected sample of students in the spring semester of their second year that had varying amounts of experience with their general education courses.

The original version of the test has since undergone multiple revisions based on item and content analyses. For example, items from the SR-6 that performed well statistically (*i.e.*, having consistently high item-total correlations over multiple administrations) were used on the SR-7. Items from previous versions of the SR that were discarded were revised and included on SR-7. Items submitted by faculty members in previous years that were not used in any previous version of the SR were also revised and added to the SR-7. More recently, an item-writing workshop for faculty members, held in 2004, resulted in the development of almost half of the SR-7 items. Item analysis of the SR-7 provided evidence that some items did not perform well. These items were removed, subsequently forming the SR-8 version of the test. Finally, during the summer of 2007, faculty from various disciplines in science and math convened and reviewed the test. Items that were construed as problematic were revised, and additional items were developed to address gaps in content coverage, leading to the current version of the test, the SR-9.

3.4. Subscores and their development

The test blueprint for the SR-9 appears in Table 1. Some items are mapped to more than one objective; thus, the number of items assessing each objective sums to a value greater than the total number of items on the test.

The first administration of the SR-9 occurred in fall 2007. The results show that internal consistency value for the overall test is reasonably high (□ = .71). However, use of subscale scores is not recommended at this time, due to lower than desirable levels of internal consistency (i.e., less than .50). These scores and other information on this pilot test can be reviewed in the Technical Information section to follow.

Table 1

Test Blueprint for SR-9

SR-9 Objectives	# Items[a]	Items
Objective A: Describe the methods of inquiry that lead to mathematical truth and scientific knowledge and be able to distinguish science from pseudo-science.	13	2, 3, 5, 10, 14, 23, 25, 26, 27, 28, 39, 40, 41
Objective B: Use theories and models as unifying principles that help us understand natural phenomena and make predictions.	7	13, 16, 17, 22, 47, 48, 49
Objective C: Recognize the interdependence of applied research, basic research, and technology, and how they affect society.	7	1, 11, 12, 30, 31, 32, 33
Objective D: Illustrate the interdependence between developments in science and social and ethical issues.	9	2, 15, 19, 20, 21, 24, 39, 40, 41
Objective E: Formulate hypotheses, identify relevant variables, and design experiments to test hypotheses.	21	3, 4, 5, 6, 7, 8, 9, 14, 18, 23, 28, 29, 34, 35, 36, 37, 38, 42, 43, 45, 46
Objective F: Evaluate the credibility, use, and misuse of scientific and mathematical information in scientific developments and public-policy issues.	13	2, 10, 19, 20, 21, 24, 25, 26, 27, 43, 44, 45, 46
Scientific Reasoning (All objectives combined)	49	1-49

[a] Some items correspond to more than one objective; therefore, the number of items assessing each objective sums to a value greater than the total number of items assessing SR.

Section 4. Administrative Procedures

4.1. Proctor qualifications and training

While administration of the SR-9 does not require intense training, proctors should be given guidance on standardized test administration. Proctor training can be accomplished in a brief session in which they are familiarized with the test instructions and the general procedures to be adhered to during the test administration. During training, proctors should be provided with the standardized instructions to be used in the actual testing session. Instructions for each mode of administration are provided in the following section.

4.2. Testing procedures

The SR-9 should be administered in a computer-based format. Clients will be sent a URL and instructions for set up. Examinees should be provided with scrap paper and a pencil. Additionally, room temperature and lighting should be appropriate for optimal testing performance. Before beginning the test, examinees should be provided with general information about the number, type and content of items on the test. Examinees should be informed of the amount of time they will be given to complete the test and what they should do upon completion of the test. It is recommended that students be given at least 45 minutes to complete the SR-9. However, if the testing time is 45 minutes and the majority of students are still working after 40 minutes, the proctor may decide to extend the testing period for another 5 minutes. When the testing time is almost over (for example, at 40 minutes), the proctor should periodically announce the time remaining (e.g. 5 minutes, 2 minutes, 1 minute).

.

4.3. Extent of exchangeability

When an instrument is administered in both paper-and-pencil and computer-based formats, information derived from one context may not be directly applicable to another context. In other words, it cannot be assumed that reliability or validity information collected through one mode of administration will generalize to the other mode (Mead & Drasgow, 1993; Wise & Plake, 1990). Equivalence between testing settings should be established before applying information collected in one setting to information collected in another.

Differences in scores were found to exist between paper-based and computer-based administrations, though the effect size was shown to be small. In general, students completing the computer-based SR-9 had no significant differences with the pencil-and-paper version. There are several factors which may influence the effect of the administration method. For example, computer-administered testing could be problematic for test takers who are unfamiliar with using a computer. In such situations, computer skills may confound the meaning of SR-9 test scores. To avoid this, it is recommended that either a trained proctor be available to assist with computer skills, detailed instructions on basic computing skills be provided, basic computer competency be a prerequisite of taking the computer based test, or some combination of these. Also, even computer savvy test-takers may encounter technical problems during a testing session that cannot be anticipated or prevented (i.e. computer freezing or power shortage). If a situation like this occurs, test scores may be unusable or biased due to frustrations or interruptions. It is recommended that a computer technician oversee the computers being used for testing and be available during testing for troubleshooting.

Besides technical problems, there are other factors to consider when selecting an administration method. For example, students taking the paper-based test could skip items (with the possibility of returning to them later), whereas students taking the computer-based test may not be able to skip items (they may be forced to make a selection to move on), or they may not be able to review previously answered items depending on how the computer-based assessment is set up. These factors may influence the comparability of scores across administration methods. Therefore, these set-up options should be carefully considered and discussed with a local computer-based testing administrator.

One clear advantage of the computer-based administration method is that test-takers cannot be unclear about their response or provide invalid (out-of-range) responses. This results in less missing data. However, in the SR-9 test, this may influence the comparability of scores obtained from the different administration methods because missing data is scored the same as incorrect data (zero).

As shown in Table 2, in general, students completing the computer-based SR-9 performed as well as those completing the paper-based form. The score differences between the two administrations were small, as indicated by the effect sizes (*d*, a standardized value, can be interpreted in terms of number of standard deviations).

Table 2

Percent-Correct Means Comparison Across Computer-Based and Paper-Based Administrations for SR-9 Scores

	computer-based				Paper-based				differences
Sample	***N***	***M***	***SD***	□	***N***	***M***	***SD***	□	**Cohen's *d***
Freshmen Fall 2007	413	47.99	7.70	.65	995	48.52	9.21	.75	.061
Sophomores Spring 2008	97	52.97	6.60	.61	923	52.72	9.12	.77	.028

Section 5. Technical Information

5.1. Scoring and interpretation

All SR-9 items are selected response. The majority of items have three response options including the correct response. The range is between two and four response options. Three response options are considered the optimal number of choices for multiple-choice test items (Rodriguez, 2005). Items are scored dichotomously: a correct response to an item is given a score of '1' and an incorrect response to an item is given a score of '0.' The total score is obtained by summing the scored item responses. Higher total scores indicate that examinees

have higher levels of scientific reasoning, and lower total scores indicate that examinees have lower levels of scientific reasoning.

5.2. Evidence of reliability

Reliability refers to the degree of stability and consistency of test scores. Due to the various sources of variability in test scores, there are different ways of measuring reliability. The SR-9 has been examined for reliability as measured by Cronbach's alpha (α), which is frequently used to determine internal consistency. Specifically, α requires only one administration and is the mathematical equivalent of the average of all possible split-half reliability computations. Alpha indicates how much variance in the observed scores is attributable to the true score. In other words, α indicates how related the scores on the items are to the construct of interest (in this case, scientific reasoning). While coefficients with a value of .70 or higher have traditionally been considered adequate for scale use, reliabilities above .80 are desirable (Nunally, 1978).

The reliability of the SR-9 test has been calculated for two administrations of the test. The reliability values for Scientific Reasoning are presented in Table 3.

Table 3

Sample Sizes, Cronbach's Alpha, Raw Scores, and Standard Deviations for SR-9

Sample	*N*	□	*M*	*SD*
Fall 2007	1408	.73	31.92	5.81
Spring 2008	1020	.76	34.81	5.88

5.3. Evidence of validity

Validity refers to the degree to which one can make inferences from the scores obtained on a test. Validity is not an absolute state, but rather a collection of evidence indicating that the scores obtained on a test are valid for their intended uses (AERA, 2000).

Content validity. Faculty from mathematics and science departments wrote the SR-9 items using the objectives as their guide, and conducted a content alignment of the SR-9 items to further verify the extent to which items were linked to the appropriate objectives. During this most recent mapping, every item in the SR-9 successfully translated back to an objective.

5.4. Norming

The SR-9 was administered to two random samples of students at JMU. The Fall 2007 sample consisted of 1,408 incoming first-year students. In Spring 2008, 1,020 sophomore students (those with 45-70 credits completed prior to assessment testing) took the SR-9. For both of these groups, the test was administered in a low-stakes environment. Under these testing conditions, students may not give their best effort; therefore, the scores may underestimate student knowledge of the intended construct. The ethnic backgrounds of the students in the sample roughly approximated those of the overall JMU population (83% white, 5% Asian, 4% Black, 2.5% Hispanic, and 5% not specified, and 0.2% other).

To determine how the students at your institution performed in relation to college students at the institution that serves as the site of SR-9 research, refer to Tables 4 and 5. Table 4 contains score information for first-year students; table 5 contains score information for college sophomores (students with 45-70 completed credits). The percentile ranks associated with each SR-9 raw score are presented for the total group and by gender.

Table 4

Percentile Ranks for SR-9 Scores for Freshmen at a Mid-Atlantic 4-Year Institution

Score	Total Group (*N* = 1408*)	Males (*n* = 530)	Females (*n* = 872)
46	99.93	99.81	---
45	99.82	99.53	--
44	99.68	99.34	99.89
43	99.22	98.58	99.60
42	98.44	97.55	98.97

41	97.37	96.51	97.88
40	95.88	94.53	96.67
39	93.43	91.23	94.72
38	89.99	87.55	91.40
37	86.19	83.77	87.56
36	81.96	78.87	83.72
35	76.88	73.40	78.84
34	70.31	66.04	72.71
33	62.86	57.83	65.65
32	55.15	50.94	57.40
31	47.87	44.72	49.48
30	41.55	38.30	43.23
29	34.98	32.36	36.24
28	28.66	26.23	29.76
27	22.98	19.53	24.66
26	18.29	15.19	19.84
25	14.74	12.83	15.65
24	11.51	10.66	11.81
23	8.70	8.49	8.66
22	6.43	7.08	5.91
21	4.79	6.04	3.96
20	3.55	4.81	2.69
19	2.56	3.58	1.83
18	1.81	2.83	1.09
17	1.31	2.26	0.63
16	0.85	1.32	0.46
15	0.50	--	0.23
14	0.28	0.57	0.06
13	--	--	--
12	0.07	0.19	--
11 or fewer	--	--	--

-- indicates that this score was not present in the data set.
* gender information was missing for 6 individuals.

Table 5

Percentile Ranks for SR-9 Scores for Sophomores at a Mid-Atlantic 4-Year Institution

Score	Total Group (N = 1020*)	Males (n = 372)	Females (n = 647)
48	99.95	99.87	--
47	99.71	99.19	--
46	99.02	97.85	99.69
45	97.79	95.56	99.07
44	96.08	92.88	97.91
43	93.68	89.92	95.83
42	90.78	86.56	93.20
41	87.01	82.80	89.41
40	81.62	77.15	84.16

39	75.49	70.56	78.28
38	68.53	63.58	71.33
37	61.52	55.91	64.68
36	54.41	48.79	57.57
35	47.21	43.15	49.46
34	40.29	37.77	41.65
33	33.87	32.53	34.54
32	28.48	28.90	28.13
31	23.77	25.40	22.72
30	19.56	21.77	18.24
29	15.20	18.01	13.60
28	11.27	14.11	9.66
27	8.58	11.42	6.96
26	6.91	9.68	5.33
25	5.49	8.47	3.79
24	4.31	7.26	2.63
23	3.68	6.18	2.24
22	3.33	5.51	2.09
21	2.99	4.84	1.93
20	2.55	4.03	1.70
19	2.01	3.23	1.31
18	1.47	2.55	0.85
17	1.08	2.02	0.54
16	0.69	1.34	0.31
15	0.29	0.67	0.08
14	0.10	0.27	--
13 or fewer	--	--	--

-- indicates that this score was not present in the data set.
* gender information was missing for 1 individual.

5.5. Meeting a standard

5.5.1. The Standard Setting Process

In 2008, 37 science and math faculty members participated in an Angoff standard setting (Angoff, 1971). This process involves having participants, or panelists, give judgments about the likelihood that a "minimally competent" student could get the item correct. For this purpose, the minimally competent student is someone who completed all of their scientific reasoning general education requirements by just passing the courses (i.e., no interest in the area beyond filling requirements, and receiving average grades in these classes). Panelists make a judgment for each item, expressed in terms of a probability carried out two decimal places (e.g. 50% probability that student would be able to complete the item correctly would be recorded as .50). The ratings for each item are then summed over the entire test; the resulting total is the panelist's cut score for the exam. After each panelist's score was computed, the median of the all the cut scores was set as the cut sore.

5.5.2. Faculty expectations

Students who have science and/or math credits when they arrive at JMU should be closer to meeting the competency or academic standard than those with no related credits. Similarly, sophomores with more coursework should do better than their peers with fewer completed courses in related areas.

Table 8 shows the faculty determined cut score and the scores achieved by students, as well as proportion of students in each category who met standards.

Table 8

Fall 2007 First-Year Student and Spring 2008 Sophomore/Junior Performance Compared with Faculty Expectation

	Faculty Standard		Freshmen with no related coursework *n*=1173	All sophomores *n*=973	Sophomores with no related coursework *n*=10	Sophomores who completed general education requirements *n*=156
SR-9 Total (49 items)	37.4	Mean	42.4	47.0	45.7	48.3
		Percent meeting standard	*13.5%*	*36.5%*	*30.0%*	*47.4%*

Section 6. Additional Information

6.1. Where to get additional information

Additional information on the SR may be obtained by contacting Madison Assessment (info@madisonassessment.com). Information may also be obtained through the following website: www.madisonassessment.com.

Section 7. References

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (2000). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Angoff, W. A. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement, 2nd edition* (pp. 508-600). Washington, DC: American Council on Education.

Dawis, R. (1987). Scale construction. *Journal of Counseling Psychology, 34*(4), 481-489.

Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized tests and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114,* 449-458.

Miller, B. J., Setzer, C., Sundre, D. L., & Zeng, X. (2007, April). Content validity: A comparison of two methods. Paper presentation to the National Council on Measurement in Education. Chicago, IL

Nunally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24* (2), 3-13.

Wise, S. L. & Plake, B. S. (1990). Computer-based testing in higher education. *Measurement and Evaluation in Counseling and Development, 23,* 3-10.

Section 8. Appendix

Notes for proctors

Students should not run any programs before or during the test. As the students arrive, please ask them to take a seat at a computer but DO NOT let them play on the computers. Verify correct student via picture id and do not let test takers use cell phones once they have entered the testing area.

Restart the computers between test sessions to clear out the computer memory.