The Quantitative Reasoning Test, Version 9 (QR-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

The Quantitative Reasoning Test, Version 9 (QR-9)

The QR-9 Test Manual

Section 1. The Nature of the Instrument

The Quantitative Reasoning Test, Version 9 (QR-9) is a 26-item multiple-choice test developed by science and mathematics university faculty. This instrument was designed to assess the quantitative reasoning skills that college students may obtain through a general education curriculum. The Quantitative 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

The QR-9 is intended to provide information about the effects of curriculum and instruction on students' learning. Therefore, test results may be used to inform curriculum and instructional improvements at the program or institution level. However, it is not appropriate to use test results to make 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 quantitative 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.

2.2. Target population

The QR-9 was designed with all undergraduate college students in mind. However, it is important to take note of the characteristics of the students in the samples that were used for test development and reliability/validity studies of the instrument. Please refer to the section on norming (section 5.4) for detailed information regarding these samples. In general, the closer the intended population to the normative sample, the closer test results will be those presented in this manual. Test users who are considering using the test on a population very different from the intended population are cautioned that test reliability and validity may not generalize.

The QR-9 is designed to be content-free, meaning that responding correctly to the items does not require specific content knowledge of any domain of science. Thus, the instrument is appropriate for students who have diverse science backgrounds and can be used to assess programs that cover various science domains.

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 QR 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

In our modern society, quantitative reasoning is an important skill. Daily tasks often require reasoning quantitatively to formulate solutions and complete everyday activities. To make informed decisions in everyday life and across a broad spectrum of occupational pursuits, students must learn to understand and interpret information in graphical or numerical forms. Thus, quantitative reasoning is considered an important component of higher education curriculum.

The QR-9 focuses on two aspects of quantitative reasoning that represent that the quantitative learning objectives of general education curriculum at JMU. One objective is the ability to use graphical, symbolic, and numerical methods to analyze, organize, and interpret natural phenomena. This is assessed through 21 items which require students to match graphical displays with numerical formula, to create and understand mathematical models, and to interpret results provided in graphical format. The second quantitative reasoning objective in the QR-9 is the ability to discriminate between association and causation and identify the types of evidence used to establish causation. This is accomplished through 10 items which require an understanding of the characteristics of a true experiment, and the limitations of conclusions that can be drawn from correlational or quasi-experimental data. Four of the items overlap and have been assigned to both objectives. The test blueprint appears in Table 1. The wordings of these objectives have been revised over time, but the main idea and scope of the objectives have been stable across revisions.

3.2. Item type selection

All QR-9 items use a selected-response format, with the number of response options ranging from two to six. 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 QR-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

The items on the QR-9 were written by science and math faculty in collaboration with assessment and measurement specialists. Open-ended interviews with representative faculty from the mathematics and science departments provided information about the objectives of the math and science component of the general education program. 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 Quantitative 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 ask 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 easily 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 general education courses.

The original version of the test has undergone multiple revisions based on item and content analyses. For example, items from the QR-6 that performed well statistically (*i.e.*, having consistently high item-total correlations over multiple administrations) were used in the QR-7. Items from previous versions of the QR that were discarded were revised and included on the QR-7. Items submitted by faculty in previous years that were not used in any previous version of the QR were also revised and added to the QR-7. More recently, an item-writing workshop for faculty held in 2004 resulted in the development of almost half of the QR-7 items. Items on the QR-8 may have been revised from their form on the QR-7 as a result of recent recommendations to reduce the number of distracters in multiple choice items to two high quality alternatives (Rodriguez, 2005). Item-analysis of the QR-7 provided evidence that some items did not perform well. These items were removed, subsequently forming the QR-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 QR-9.

3.4. Subscales and their development

The basis of subscales for the QR-9 is a series of back-translation exercises which resulted in grouping items by two main learning objectives. In the back translations exercises, subject-matter experts (i.e. the faculty teaching the curriculum) indicated which items they felt were functioning as assessments of which curricular objective. Item mappings were combined from all experts and group discussion took place about contradictory mappings. The test blueprint presented in Table 1 resulted from this process.

Although technically the QR-9 items can be grouped according to these two objectives, and the objectives have the content validity of several rounds of back translations, it is not recommended that the subscale scores for the objectives be used because of the lower-than-desirable values of Cronbach's alpha reliability estimates. Thus, only total scores for the test are reported in this manual. Test administrators interested in using subscales scores may wish to add items to the subscales and re-test the appropriateness of using the lengthened subscales.

Table 1

Test Blueprint for QR-9

Scales	# of Items[a]	Item numbers
Quantitative Reasoning (total test)	26	1-26
Use graphical, symbolic, and numerical methods to analyze, organize, and interpret natural phenomena.	21	2-13, 18-26
Discriminate between association and causation, and identify the types of evidence used to establish causation.	10	1, 14-17, 20, 23-26

[a] Some items correspond to both objectives; therefore, the sum of number of items assessing each objective is greater than the total number of items assessing the QR subscale.

Section 4. Administrative Procedures

4.1. Proctor qualifications and training

While administration of the QR-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 QR-9 test should be administered in 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 25 minutes to complete the QR-9. However, if the majority of students are still working after 20 minutes, the proctor may decide to extend the testing period for another 5 minutes. When the testing time is almost over (for example, after 20 minutes), the proctor should periodically announce the time remaining (e.g. 5 minutes, 3 minutes).

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 should be established before applying information collected in one setting to information collected in another.

Differences in scores have been 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 QR-9 did not perform as well as those completing 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 QR-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, or basic computer competency be a prerequisite of taking the computer based test. 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 QR-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).

Table 2 presents results of the QR-9 means for two samples of students (freshmen and sophomores) completing the two administration methods. The data indicate that freshmen students completing the computer-based test performed less well. Score differences may have occurred because of differences in administration procedures between paper and computer based administrations (see descriptions above). Given that the test is administered in low-stakes conditions, examinee motivation can affect student scores. The computer-based administration procedures allow students to leave as soon as they have completed the test, unlike the paper-based administration which requires students to wait until all students are finished. Allowing students to leave when finished may encourage rushing through the test when scores do not have consequences for students. This possibility will be explored further in future administrations of the test. Alternatively, score differences may have occurred because of differences in the way the items are presented on computer versus on paper. Particularly, some items require scrolling, which may be distracting for students. Also, some items that are part of a set on the paper-based exam are presented on separate screens on the computer-based exam. Again, these potential sources of variability are currently being explored.

Table 2

Percent-Correct Means, Standard Deviations and Reliabilities of Computer-Based and Paper-Based Administrations of QR-9

	Computer-based				Paper-based				Differences
Sample	*N*	*M*	*SD*	□	*N*	*M*	*SD*	□	*d*
Freshmen Fall 2007	413	60.23	13.85	.60	995	62.56	14.52	.65	.28
Sophomores Spring 2008	97	67.80	12.77	.57	923	67.47	14.54	.67	.02

Section 5. Technical Information

5.1. Scoring and interpretation

All QR-9 items are selected response. The majority of items have three response options including the correct response. The range is between two and six 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 quantitative reasoning, and lower total scores indicate that examinees have lower levels of quantitative reasoning.

Administrators using this instrument are encouraged to aggregate scores across at least 30 students prior to making inferences about a population or program.

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 QR-9 has been examined for reliability as measured by Cronbach's alpha (α), which is frequently used to determine internal consistency. Specifically, Cronbach's alpha 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, alpha indicates how related the scores on the items are to the construct of interest (in this case, quantitative 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 QR-9 test has been calculated for two administrations at JMU. The reliability values for the QR-9 are presented in Table 3.

Table 3

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

Sample	*N*	□	*M*	*SD*
Fall 2007	1408	.64	16.09	3.73
Spring 2008	1020	.66	17.55	3.74

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 QR-9 items using the objectives as their guide, and conducted a content alignment of the QR-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 QR-9 successfully translated back to an objective.

5.4. Norming

The QR-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 theQR-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 QR-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 QR-9 raw score are presented for the total group and by gender.

Table 4

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

Score	Total Group (*N* = 1408*)	Males (*n* = 530)	Females (*n* = 872)
26	--	--	--
25	99.82	99.62	99.94
24	99.11	98.21	99.66
23	97.51	95.38	98.80
22	94.53	90.94	96.67
21	90.59	85.75	93.46
20	85.05	78.49	88.93
19	76.46	67.55	81.71
18	66.97	56.70	72.99
17	57.63	47.08	63.88
16	48.08	38.58	53.78
15	38.64	30.66	43.41
14	29.19	22.64	33.03
13	21.06	16.04	23.91
12	14.67	10.75	16.86
11	9.66	7.26	10.95
10	5.75	4.91	6.14
9	3.23	3.40	3.10
8	1.92	2.45	1.61
7	0.85	1.32	0.57
6	0.25	0.57	0.06
5	--	--	--
4	0.07	0.19	--
3	--	--	--
2	--	--	--
1	--	--	--

Note: Based on those freshmen who took the test in Fall 2007 administration.
-- indicates that this score was not present in the data set.
* Gender information was missing for 6 individuals; these students received scores on the QR-9.

Table 5

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

Score	Total Group (*N* = 1020*)	Males (*n* = 372)	Females (*n* = 647)
26	--	--	--
25	99.12	98.39	99.54
24	97.21	95.16	98.38
23	93.87	89.65	96.29
22	88.97	82.80	92.50
21	81.86	74.73	85.94
20	72.94	65.32	77.28
19	62.75	55.11	67.08
18	52.35	45.70	56.11
17	41.72	35.75	45.05
16	31.62	26.48	34.47
15	23.28	20.03	25.04
14	16.23	14.78	16.92
13	11.27	11.16	11.21
12	7.94	8.60	7.50
11	5.34	6.32	4.79
10	3.33	4.17	2.86
9	1.96	2.55	1.62
8	1.27	1.75	1.00
7	0.83	1.21	0.62
6	0.59	0.94	0.39
5	0.39	0.67	0.23
4	0.20	0.40	0.08
3	0.05	0.13	--
2	--	--	--
1	--	--	--

Note: Based on those sophomores who took the test in Spring 2008 administration.
-- indicates that this score was not present in the data set.
* Gender information was missing for 1 individual; this student received a score on the QR-9.

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 quantitative and 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 their higher ed school 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
QR-9 Total (26 items)	19.4	Mean	15.8	17.6	18.0	18.2
		Percent meeting standard	*15.4%*	*32.3%*	*50.0%*	*43.6%*

Section 6. Additional Information

6.1. Where to get additional information

Additional information on the QR–9 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