Non-standardized tests: psychometric analysis

D. Charkova

Plovdiv University "Paisii Hilendarski", Plovdiv, Bulgaria
Corresponding author. E-mail: dcharkova@gmail.com

Paper received 25.08.2016; Accepted for publication 05.09.2016.

Abstract. This article examines the pedagogical application of the statistical software jMetrik as a tool that can be used by teachers to ascertain the reliability and quality of the tests they create to assess their students' achievement. The article offers an explanation of key terms in classical test theory, including coefficients of difficulty and discrimination, distractor efficacy, and reliability coefficients. This is followed by a step by step introduction to jMetrik and its use for conducting psychometric analysis. The process is illustrated through sample data, screen shots, and relevant examples.

Keywords: *classical test theory, item analysis, multiple-choice questions, distractor efficacy, reliability coefficients, norm-referenced testing.*

Introduction. According to Brown [1], tests can be categorized in different ways according to their purpose and philosophical approach. For example, diagnostic, placement, and achievement tests serve a different purpose. On the other hand, norm-referenced and criterion-referenced tests are based on two different approaches to testing. In norm-referenced tests, students' scores are ranked and compared to each other before grades are assigned. Following the properties of a normal distribution only a small percentage of students (approximately 16%) get higher grades, whereas the majority of the students receive average grades (64%). Likewise, approximately 16% of the students get lower grades. Criterion-referenced testing follows a very different approach, where students' grades are assigned in comparison to a *priori* set criterion [7]. In this approach, students' scores are not compared to each other, but to the criterion, and grades are assigned according to whether the students have met the criterion or not, and how close and far from the criterion their scores are.

From a completely different point of view, which is not often mentioned in testing literature, educational tests can be categorized into standardized and teacher developed tests. Standardized tests are normally developed by a team of specialists in testing, subject matter, and statistics, who put painstaking efforts into developing, pilot testing, and improving the test items. These specialists have all the know-how they need to create valid and reliable tests.

In contrast, it is still a very common occurrence to have teachers develop their own tests without any or very little training in test design and test statistics. This article is, therefore, not written for use by testing specialist, but for teachers who find themselves developing tests without knowing whether the tests they have developed can be trusted as reliable measures of students' achievement. Further, this article is also meant for those teachers who follow a norm-referenced approach to grading their students because the statistics that will be computed and discussed are of little relevance to criterion-referenced testing [1], [7].

What statistics would help a teacher evaluate the quality of a test? First of all, it should be made clear that the statistics that will be discussed in this section are calculated only after the test is administered. So, they complement what a teacher should do before and during the test to ascertain the quality of a test. For instance, in the development of a multiple-choice test, teachers need to observe the principle of validity as well as the standards of writing good stems and options [2]. During the test, they need to guarantee optimum conditions for students to work on their test to rule out cheating. While the pre and during stages of creating and administering a test are very important, they are beyond the scope of this article, which focuses on what teachers can do after the test has been administered and scored in order to find whether all items in the test are effectively constructed and correctly scored.

Statistics that attest to the quality of a test item.

There are three main types of statistics that can be very useful to teachers in evaluating the items on the tests they have administered to the students: Coefficient of Item Difficulty (p); Coefficient of Discrimination (ID) and Distractor Efficacy.

***Coefficient of Item Difficulty* (p):** This statistics shows the level of difficulty of a test item: easy, moderate, or difficult. It is calculated by the dividing the total number of students who chose the correct answer for a given item by the total number of students who took the test. The difficulty coefficients vary between 0 and 1. High p-values show that the items are easy; low p-values show that the items are difficult. For example, *p*-values between .9 to 1 = very easy items; *p*-values below .20 = very difficult. Shrock and Coscarelli [7] advise teachers who follow the norm-reference approach to include items with *p*-values ranging between .30 and .70, and to exclude items outside this range. However, it should be kept in mind that this is just a recommendation which may be applied with some flexibility. For example, sometimes it may be reasonable to keep an easy item if it covers a concept/knowledge that every student should know, or a difficult item if the item corresponds to the objective, has no flaws, and is missed by the low scoring students but answered correctly by the high scoring students.

***Coefficient of Discrimination (ID)*:** The coefficient of item discrimination is one of the most useful statistics because it shows how well an item discriminates between high scorers on the test and low scorers on the same test. There are different ways to calculate the discrimination index. For items which are scored dichotomously (1= correct answer; 0 = wrong answer), the point- biserial correlation is equal to the Coefficient of Discrimination [3]. Shrock and Coscarelli [7] explain that the point biserial coefficient shows the correlation between test-takers' performance on a single item with their total score on the test. The values can range between -1 and + 1. A positive coefficient shows that the students who did well on the test also did well on the specific item. In contrast, a negative value signals that the students who earned a low score on the whole test, did well on this specific items. Such items should be removed from the test because they do not discriminate between the high scoring and low scoring students in a reliable way. Different authors give slightly different reference values, but as a general rule: DI = .40 and higher shows very good discrimination power; DI = .30 to .39 indicates good discrimination power; DI = .20 to .29 = acceptable discrimination power; DI = .19 and below = poor discrimination power; DI = 0 shows no discrimination; DI = negative shows the item has serious flaws. Usually items with poor or negative discrimination power tend to have at least one of the following problems: a controversial correct answer; more than one correct answer; no real correct answer; an ambiguous stem; a wrong key.

Of the two item statistics, p = difficulty and DI = discrimination power, the more important one is the DI. It is indicative of problematic items more than p. Usually there is no real connection between the two indexes, except in the case of items which are very easy and very difficult. Such items usually have low DI, that is they do not discriminate between high and low scoring students.

Distractor efficacy: Another statistic that can help teachers improve test items is the distractor efficacy which attests to whether a distractor does what it is supposed to do – appear a plausible correct answer to the students who do not know. The distractor efficacy is conveyed through the same statistics: the Coefficient of Difficulty (p) and the Discrimination Index (ID, however, they are interpreted in the opposite way to the one for test items. In the case of distractors, *p* should be lower than the one for the item, otherwise, it would indicate that the majority of the students selected one particular distractor rather than the correct answer. Also, in the best possible scenario, the p-values of the distractors will be similar, which will indicate that all distractors are attracting students who do not know, not just one of them. An example of good distractor efficacy as shown by the p-values will be the following: correct choice: p = .54; distractor 1: p = .19; distractor 2: p = .15; distractor 3: p = .12. The difficulty coefficient of the correct answer is the highest and shows an item of moderate difficulty, whereas the other 46% are distributed in similar proportions among the distractors.

The DI or the point-biserial correlation for the distractors shows good efficacy if it has negative or very low values, below .19. Specifically, it indicates that the students who had overall low total scores on the test, chose the distractors with negative or low values. An example of good distractor efficacy are the following discrimination values: correct answer = .57; distractor 1 = -.40; distractor 2 = -.26; distractor 3 = -.13.

***Coefficient of internal consistency*:** According Webb, Shavelson, and Haertel [6], another important indicator for the quality of test items are the reliability coefficients. There are several different ways to calculate reliability, named after the statisticians who discovered these methods. Some of them include: Cronbach's Alpha, Guttman's L2, Feldt-Gilmer, Feldt-Brennan, and Raju's Beta. Each of these coefficients can take a value between 0 and 1, where 0 shows no consistency among the items on the test, and 1 indicates very high consistency. However, a very high coefficient of constancy, above .9, shows that the items are repetitious or that there are more items than necessary [5]. Of all the above coefficients, Cronbach's Alpha is the most widely used for estimating test-score reliability from a single test administration, based on the relationship among test items.

Whichever of the coefficients is used, the interpretation is similar. A test with a good internal consistency should have a coefficient between .75 to .95. This means that 75% to 95% of the items on the test should consistently measure the same construct/content. In addition, the reliability statistics also show how the coefficient will change if a specific item is deleted. Normally, if the overall coefficient increases after the deletion of an item, this indicates that the item does not fit well with the other items. That is, it may test content which is different from the content tested by the rest of the items. However, a small increase of the overall coefficient is not a good reason to delete an item from a test. For example, if Coefficient Alpha for the whole test = .87 and after deleting Item 8, Alpha changes to .88, an increment by one hundredth is not a sufficient increase to delete Item 8 from the test.

As a rule of thumb, the overall coefficient of internal reliability and the change in its value after deleting each one of the items should be considered in relation to the rest of the item statistics (p, DI, and distractor efficacy) in order to identify poor items, which, then, should be reexamined for problems in the stems, options, key, and content, before deciding whether they should be deleted.

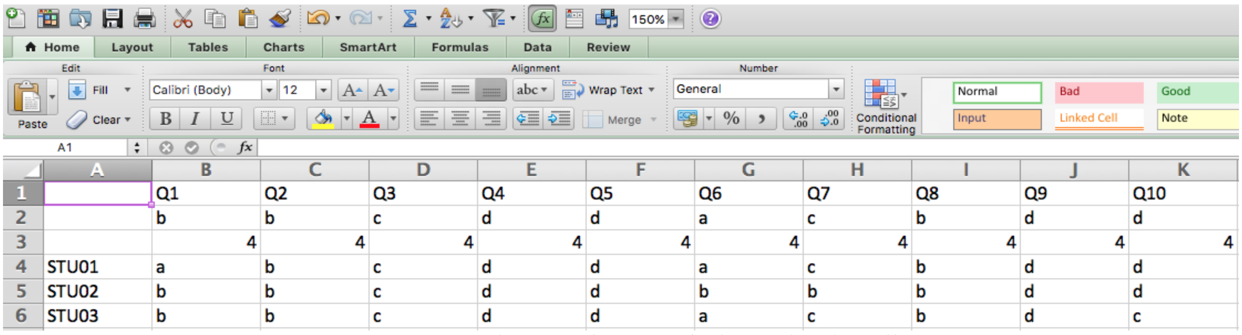
jMetrik: a tool for computing item statistics.

According to the jMetrik homepage [4], jMetrik is a free and open source computer program for psychometric analysis (http://www.itemanalysis.com/). The fact that it is free and fairly easy to use, makes it an ideal tool for teachers who want to improve the quality of the tests through item analysis.

Methodology. To illustrate the use of jMetrik, we will take a small sample data, elicited through a 20-item multiple choice test, administered to students in an Information Technology program of study. The reader should be reminded that the small sample size is used for the sake of brevity, however, in a real testing situation, the number of items on the test should cover the learning objectives and would normally be much higher [7]. Each question had 4 options, among which one was the correct answer and the remaining three were distractors.

In the remaining part of this article, we will use the sample test data to illustrate the following functions of jMetrik: 1) Downloading jMetrik; 2) Preparing an .xlsx file with test data; 3) Creating a new data base in jMetrik; 4) Importing and preparing your test data for analysis with jMetrik; 5) Calculating and interpreting the coefficients of difficulty and discrimination; 6) Performing and interpreting item reliability analysis; 7) Using the statistics to evaluate and improve your test items.

Step by step introduction to jMetrik. ***Step 1***: Download jMetrik from http://www.itemanalysis.com/ /jmetrik-download.php. Choose the version suitable for your computer (options for Windows and Mac OS are available). ***Step 2*** contains three sub steps to prepare the .xlsx table to be imported into jMetrik. The first includes transforming all answers into multiple choice options (a, b, c or d, if they are 4 options) as shown in the screen shot in Fugure 1. Second, insert two rows at the top of your table. The first row (row 2, Figure 1) should have all correct answers from the test; the second one (row 3, Figure 1) the number of multiple choice options available for each item. For example, if there are 4 options (a, b, c, and d) enter the number 4. When this is completed as illustrated below, the .xlsx document should be saved as a .csv file, otherwise jMetrik will not allow it to be imported.



	A	B	C	D	E	F	G	H	I	J	K
1		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
2		b	b	c	d	d	a	c	b	d	d
3		4	4	4	4	4	4	4	4	4	4
4	STU01	a	b	c	d	d	a	c	b	d	d
5	STU02	b	b	c	d	d	b	b	b	d	d
6	STU03	b	b	c	d	d	a	c	b	d	c

Figure 1. Preparing your document for import into jMetrik

In ***Step 3*** create a new database as shown in Figure 2 below. Go to Mange > New Database and name it accordingly. Then go to Manage > Open Database and open the newly created database.

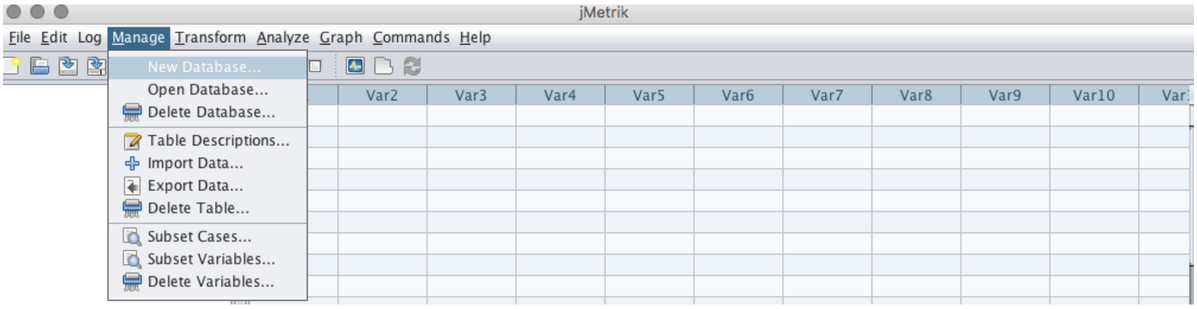
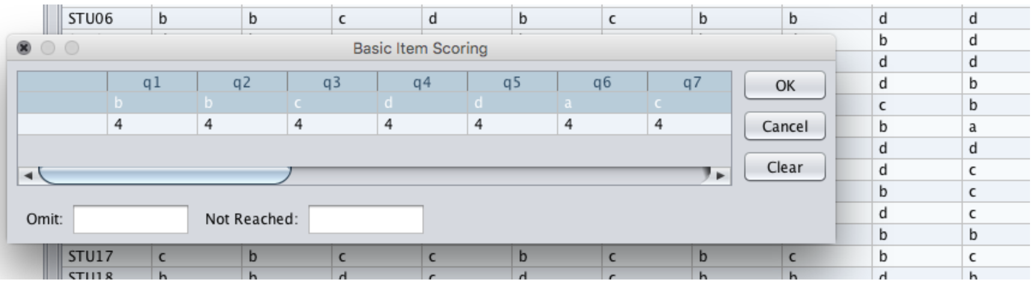


Figure 2. Creating a database in jMetrik

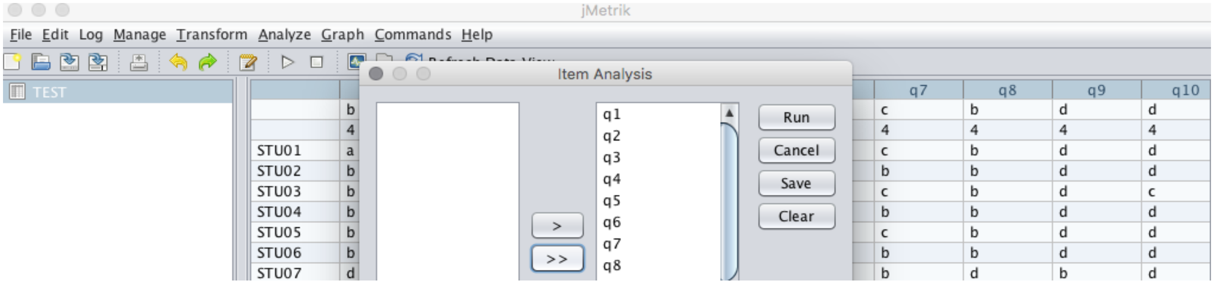
In ***Step 4*** import the .csv file into jMetrik. Go to Manage> Import Data. Select your file, name it accordingly and click the Import button. In this step there is one more specificity. Go to Transform> Basic Item Scoring. A Basic Item Scoring dialog box will appear as shown in Figure 3. Enter all correct answers as well as the number of possible answers for every item in the test. This is analogical to the actions in Step 2 (cf. Figure 3). Then click OK. A table with analysis of all test items will appear.



	q1	q2	q3	q4	q5	q6	q7
	b	b	c	d	d	a	c
	4	4	4	4	4	4	4

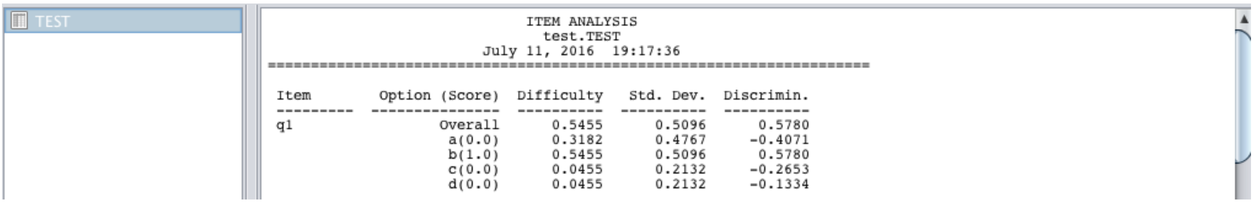
Figure 3. Basic Item Scoring: entering answers and multiple choice options

Step 5: After having completed all the previous steps, the analysis of the test data can begin. To obtain the Coefficient of difficulty *p* and the Coefficient of discrimination (*ID*), go to Analyze > Item Analysis. Select all test questions and move them to the right side of the window, as shown in Figure 4. When done, click Run. In order to illustrate how to use the calculated *p* and *ID* values, we will explore item 1 (q1) in Figure 5. From the information provided in the second column, we can see that the correct answer is *b* (it has a score of 1 and all the rest have 0). More than half of the students (54%) have chosen the right answer. If the remaining percent were distributed evenly among the other three distractors, this question would not be problematic. However, option *a* has attracted a lot more students than options *c* and *d*. This means that choices *c* and *d* must be revisited and re-examined in order to decide whether they need to be revised.



		q7	q8	q9	q10
	b	c	b	d	d
	4	4	4	4	4
STU01	a	c	b	d	d
STU02	b	b	b	d	d
STU03	b	c	b	d	c
STU04	b	b	b	d	d
STU05	b	c	b	d	d
STU06	b	b	b	d	d
STU07	d	b	d	b	d

Figure 4. Selecting test items for analysis

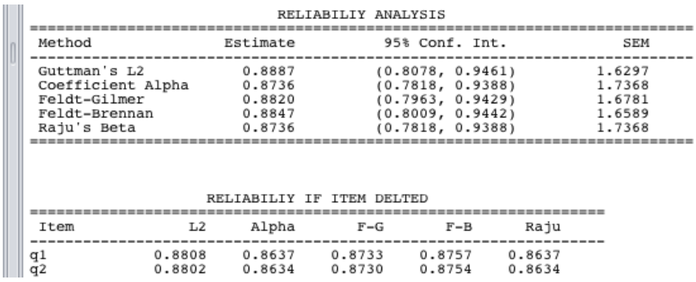


```
                         ITEM ANALYSIS
                           test.TEST
                     July 11, 2016  19:17:36
=====================================================================

 Item        Option (Score)  Difficulty   Std. Dev.   Discrimin.
 ---------  ---------------  ----------  ----------  ----------
 q1                 Overall      0.5455      0.5096      0.5780
                     a(0.0)      0.3182      0.4767     -0.4071
                     b(1.0)      0.5455      0.5096      0.5780
                     c(0.0)      0.0455      0.2132     -0.2653
                     d(0.0)      0.0455      0.2132     -0.1334
```

Figure 5. Analyzing Coefficient of Difficulty and Coefficient of Discrimination

In ***Step 6***, Reliability Analysis is performed in order to find out how each item influences the entire test if deleted. As discussed in the theoretical part, there are several different methods to calculate reliability, see Figure 6. Since Cronbach's Alpha is the most commonly used formula we will use it for reference (Coefficient Alpha = 0.8736). When we look at q1 and compare its Alpha to 0.8736 we see that it decreases. This mean that if q1 were to be deleted then the overall reliability of the test would also decrease.



```
                         RELIABILIY ANALYSIS
============================================================================
 Method                 Estimate          95% Conf. Int.              SEM
----------------------------------------------------------------------------
 Guttman's L2             0.8887        (0.8078, 0.9461)            1.6297
 Coefficient Alpha        0.8736        (0.7818, 0.9388)            1.7368
 Feldt-Gilmer             0.8820        (0.7963, 0.9429)            1.6781
 Feldt-Brennan            0.8847        (0.8009, 0.9442)            1.6589
 Raju's Beta              0.8736        (0.7818, 0.9388)            1.7368
============================================================================


                    RELIABILIY IF ITEM DELTED
==================================================================
 Item             L2      Alpha        F-G        F-B       Raju
------------------------------------------------------------------
q1            0.8808     0.8637     0.8733     0.8757     0.8637
q2            0.8802     0.8634     0.8730     0.8754     0.8634
```

Figure 6. Reliability statistics

Step 7: Considering the test statistics discussed in the introduction of this article and the results of the psychometric analysis, scrutinize each items on the test and identify the problematic ones. First of all, look at the column showing the discrimination index. Any item with a negative or very low discrimination power should be of concern. Such items are likely to have low distractor efficacy too. Such items will most probably need to be removed from the test. Next, look at the remaining items, check their *p* values, distractor efficacy, and reliability if items deleted. This will provide ground for making an informed judgement about items that need to be deleted or revised. The best way to make these decisions is to revisit the items that were brought to attention by the statistics, look at their stems, distractors, and the learning objectives they are supposed to measure. Sometimes, a fairly easy item may turn out to be important for measuring knowledge of a basic learning objective, so it may be retained it the test. Remember that statistics alone do not make sense. They need to be interpreted in the context of the subject matter and the learning goals.

Conclusion. The reader should be reminded that this has a limited scope as it focuses on the test statistics after the test is administered and scored. The article does not cover the different types of validity and reliability that should be observed in the process of writing and administering the test. This article is mostly relevant to teachers who develop their own tests and are interested in using psychometric tools for improving the test items.

Acknowledgments: This publication has been partially sponsored by grant Plovdiv University, contract number MU15-FMIIT-008.

REFERENCES

1. Brown, J. (1996). *Testing in language classrooms*. New Jersey: Prentice Hall Regents.
2. Brown, H.D., & Abeywickrama, P. (2010). *Language assessment*: *Principles and classroom practices*. White Plains, NY: Pearson Education.
3. Glass, G. V., & Hopkin, K. D. (1996). *Statistical methods in education and psychology (3rd ed.).* Boston: Allyn & Bacon.
4. jMetrik (2007-2015). Psychomeasurement Systems, LLC. Retrieved on August 18th, 2016 from http://www.jmetrik.com/index.php

5. Morgan, G.A., Leech, N.L., Gloeckner, G.W., Barrett, K. C. (2013). IBM SPSS for introductory statistics (5th edition). New York: Taylor and Francis.
6. Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2007). Reliability coefficients and generalizability theory. In C.R. Rao and S. Sinharay (eds.), *Handbook of statistics, Vol. 26, Psychometrics.* Amsterdam, the Netherlands, Elsevier Science B.V., 81- 124.
7. Shrock, S.A., & Coscarelli, W.C.C. (2007). *Criterion-referenced test development: Technical and legal guidelines for corporate training (3rd ed.).* Alexandria, VA: International Society of Performance Improvement.

Нестандартизированные тесты: психометрический анализ

Д. Шаркова

Аннотация. Настоящая статья рассматривает педагогическое приложение статистического софтуера jMetrik в качестве инструмента преподавателя в целях измерения надеждности и качества тестов, которых преподаватели используют для проверки знаний и достижений. Статья предлагает дефиниции ключевой терминологии классической теории тестов, которые включают коэфициенты трудности и дискриминации, эффективность дистракторов и коэффициент надеждности. Кроме того в статье рассмотрены детальное введение в программу jMetrik и ее приложение в психометрическом анализе теста. Процесс проилюстрирован примерными данными, фотографиями и подходящими примерами.

Ключевые слова: *классическая теория тестов, анализ примерного элемента теста, вопросы с множественным ответом, эффективность дистракторов, коэффициент надеждности, нормированное проведение тестов.*