

EXAMINING THE PROPERTIES OF THE STANDARD COMPREHENSIVE
EXAMINATION THROUGH THE USE OF FACTOR ANALYSIS AND
MULTIDIMENSIONAL SCALING

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE
IN THE GRADUATE SCHOOL OF THE
TEXAS WOMAN'S UNIVERSITY

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
COLLEGE OF ARTS AND SCIENCES

BY

AMJAD ALMALAQ, B.S.

DENTON, TEXAS

MAY 2016

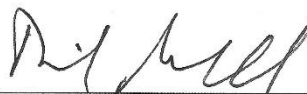
TEXAS WOMAN'S UNIVERSITY

DENTON, TEXAS

December 14, 2015

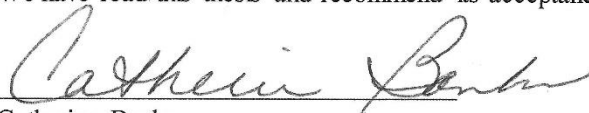
To the Dean of the Graduate School:

I am submitting herewith a thesis written by Amjad Almalaq entitled "Examining the Properties of the Standard Comprehensive Examination Through the use of Factor Analysis and Multidimensional Scaling." I have examined this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science with a major in Mathematics.



Dr. David D. Marshall, Major Professor

We have read this thesis and recommend its acceptance:



Catherine Banks



Dr. Donald Edwards



Department Chair

Accepted:



Dean of the Graduate School

ABSTRACT

AMJAD ALMALAQ

EXAMINING THE PROPERTIES OF THE STANDARD COMPREHENSIVE EXAMINATION THROUGH THE USE OF FACTOR ANALYSIS AND MULTIDIMENSIONAL SCALING

MAY 2016

The purpose of this study was to examine the properties of a departmental comprehensive examination in the course of Elementary Statistics-I (MATH 1703) as used by the Department of Mathematics & Computer Science in Texas Woman's University during the semesters Fall 2012 through Spring 2015. Item performance was assessed with standard discrimination and difficulty indices. Factor analysis and multidimensional scaling were used to assess construct validity from unique perspectives. Internal consistency reliability was defined with Cronbach's alpha coefficient. The data and findings support use of the examination, while studies of predictive and concurrent validities remain to be done.

TABLE OF CONTENTS

	Page
ABSTRACT	iii
TABLE OF CONTENTS	iv
LIST OF TABLES	v
LIST OF FIGURES	vi
 Chapter	
I. INTRODUCTION	1
Overall Strategy	2
II. RATIONALE AND THE NEED FOR THE STUDY	4
III. MATHEMATICS APPLIED	7
Factor Analysis	7
Multidimensional Scaling	10
IV. RESULTS	13
Reliability Analysis	19
Item Analysis	21
Distractor Analysis and Item Difficulty	22
V. CONCLUSION	25
VI. REFERENCES	27

LIST OF TABLES

Table	Page
1 Varimax Rotation of Components	14
2 MDS Coordinate Values	16
3 Alpha Reliability for Examination.....	20
4 Discrimination of Items	22
5 Item Performance: Distractors and Difficulty, Form A	23
6 Item Performance: Distractors and Difficulty, Form B	24

LIST OF FIGURES

Figure		Page
1	Profile plots for two dimensions of test items	16
2	MDS of the exam items	18

CHAPTER I

INTRODUCTION

Nearly 50 years ago Norman Gronlund identified characteristics of classroom achievement tests which are in strict agreement with modern criteria for assessing the effectiveness of classroom instruction and accreditation standards (Gronlund, 1968). Specifically, tests should measure clearly defined learning outcomes that coordinate with overall objectives, and tests should be useful for improving student learning. Related to the mission of using tests in the educational process is the assessment of the test itself in terms of how well the test problems function in assessing relevant learning, and the extent to which we can trust a test to provide useful and accurate information about student learning.

The present thesis considers one form of achievement test, the multiple-choice format, in assessing statistics knowledge and skills. In the end, studies on a classroom achievement test should determine the efficacy of the test, aiming at demonstrating whether and to what extent a test yields “dependable” data. Magnussen’s classic statement says it best: “Data should thus be dependable from two points of view--they should be meaningful and they should be reproducible” (Magnussen, 1966, pg. 59). The meaningfulness of a test score depends on the measured “validity” of the test; a valid test provides scores that actually convey information that the teacher or administrator needs to know about the student; or as it is commonly put, the test measures what it is supposed to measure. Reproducibility of scores reflects the accuracy or “reliability” of the measure; a reliable test measures a skill or knowledge

with precision. Dependable tests are thus valid and reliable. The current study assessed certain validities and reliabilities of the departmental final exam used in an elementary statistics, core curriculum course.

Overall Strategy

A large number of current studies, in both education and psychology, have dealt with test score profiles (Hill, Sleep, Lewis, & Ball, 2007). The term "profile" has been utilized as a part of the instruction settings and score reporting in mathematics, reading comprehension, and basic intuition aptitudes (Johannsdottir, 2013; Schmidt, Wang, & McKnight, 2005). It is not unusual for undergraduates to get test results as score profiles, speaking to the quality and shortcomings in their performance on tests (Jacobs et al., 2006; NCTM, 2000). Because of this basic practice in instruction, the profile examination was considered by training professionals as portraying basic test scores. Therefore, a mixture of exploratory systems has been utilized to recognize profile designs in mathematics. While the object of this thesis is the quality of one classroom test, multivariate procedures were applied to the test items with a view to determining if “performance profiles” might exist within the test, among the test items.

Statistics and statistical mathematical models are important tools used in conducting research and making inferences that can be applied to whole populations (Mewborn, 2003). However, despite their popularity and feasibility for use as scientific tools for accepting or rejecting hypotheses in research, they have some limitations and shortfalls (Bechger & Maris, 2015). For example, statistics has limitations in its ability to answer research questions that can be applied to the general

population. Furthermore, people are limited in their ability to understand and fully comprehend statistics, resulting in a reduction in their applicability in real world problems (Petroski, 2005). Understanding the problems and pitfalls in using certain statistical models and mathematics in different situations can help us know what models should be used in particular situations (Macnab & Payne, 2003; Mewborn, 2003). From this perspective and through this study, I aim to examine how properties of the standard comprehensive examination may help us to identify criteria that can lead to the improvement of such examinations, in a language that may be more easily understood.

CHAPTER II

RATIONALE AND THE NEED FOR THE STUDY

Examination of the properties of standard comprehensive examinations has been emphasized as part of a standards-based conception in contemporary mathematics education (NCTM, 2000). Validation of a test involves many sources of information gathered over time (Kane, 1992). However, the results of validity investigations may be affected by sources of error. Generally, statistical methods are just true on average because statistics and statistical mathematical models represent expectations about aggregate facts. Single observations do not constitute irrefutable statistics, and statistical results can be misleading. Using statistics to make decisions might result in errors, especially in inferential statistics, because it is impossible to know whether an error has been committed or the converse is true.

Regardless of cautions about the use of methods that provide information that is not completely “truthful” because of error involved, this thesis used several methods to make estimations of test properties, knowing that the estimations would carry some error. Multidimensional scaling (MDS) and factor analysis are among the common statistical mathematical models. In fact, MDS is considered a factor analysis alternative. They are appropriate for analysis of the interpretation and representation of complex structural data (Borg, Groenen, & Mair, 2013; Groenen & Borg, 2014). The aim of analyses is to unearth meaningful underlying aspects that make it possible for the researcher to explain dissimilarities or similarities observed in objects they have investigated (Borg, Groenen, & Mair, 2013; Groenen & Borg, 2014). Moreover,

such analysis helps to improve the mathematical thinking and reasoning of learners in the field of mathematics.

With regard to factor analysis, similarities in variables are shown in correlation matrices, while for MDS, as well as correlation matrices, any kinds of dissimilarity or similarity matrices can be analyzed (Jaworska & Chupetlovska-Anastasova, 2009). For instance, age and height have a probabilistic distribution in humans with a stochastic relation; knowing the age of a person can influence the determination of the probability of the person being over six feet tall. This relationship can be formalized using a linear regression model, but the regression model must include an error term, implying that age can be used to predict height, with a certain level of error.

A model that is admissible, using the example of age and height, has to be consistent with all the data points; therefore, a relation such as $\text{Height } h_j = I_j + b_1 \text{Age}_j + E_j$ is not a suitable model for the data because it cannot fit all data points exactly where all data points lie perfectly in a straight line (Borg, Groenen, & Mair, 2013; Groenen & Borg, 2014). That is why we must include the E (error value) within the model to make the model consistent with all data points, so making an inference would have to entail making assumptions about the error factor's probability distribution. For example, the error distributions in the formula can be assumed to be Gaussian, having a mean of zero. For these reasons, it is important and highly significant that the statistical mathematical models and their assumptions reflected in data be evaluated using scientific research approaches. Such attention to detail will help professionals and researchers to develop better research models using statistics to

reduce the errors associated with using statistical models to enhance research inferences.

CHAPTER III

MATHEMATICS APPLIED

Factor analysis and multidimensional scaling are the methodologies of this study. Since factor (component) analysis and multidimensional scaling are the major procedures which are applied to the exam data, they, as treatments of those statistics, are the mathematical bases for this thesis. In the following sections, I review the two procedures, factor analysis and multidimensional scaling, which are applied in this study and will be used as the methodology of the research.

Factor Analysis

With the advancement of technology, factor analysis has become more accessible in a wider variety of technological formats. Programs have been developed that emphasize the use of factor analysis, such as BMD, DATATEXT, OSTRIS, SAS, and SPSS (Kim & Mueller, 1978a). Despite the wide availability of programs capable of conducting factor analyses, the user may not have the necessary expertise to analyze the data relating to or the results from factor analysis. This lack of expertise does not stop many researchers from utilizing factor analysis in their own research, even if it yields inaccurate results.

To begin with, it is assumed in factor analysis that measured variables are linear with underlying variables. Thus, both types of variables are assumed to be in existence. However, the logical properties of factor analysis suggest that correspondence between variables occurs in that the causal variable system always leads to correlated measured variables. At the same time, the measured variable

system does not always lead to correlated causal variables (Kim & Muller, 1978b).

This establishes limited conditions available to ascertain the causal system in relation to the correlation of measured variables. This effect occurs through inference and other uncertainties of statistics. Therefore, factor analysis is affected by inferences and uncertainties of statistics.

Factor analysis is defined as a group of statistical techniques for the purpose of achieving a common objective – the representation of set variables in relation to hypothetical variables. Thus, factor analysis also focuses on interrelationships between variables. In one example, when obtaining opinions from, hypothetically, one thousand random participants, the measured variable would be the opinions provided by the participants, while the underlying variable may be a correlation coefficient. This would allow researchers to ascertain relationships among the measured variables (Kim & Muller, 1978a). Once the relationships are established, factor analysis can be used to determine if correlations occur due to other variables that are associated with the measured variable.

The analysis of these correlations and relationships allows researchers to break down complicated theories into specific steps in order to make up the whole in context of the overall theory. At the same time, the underlying variable may be unknown, prompting the researcher to utilize factor analysis to ascertain this variable for comparison. As a result, the third use of factor analysis is exploratory. In other cases, factor analysis is used to confirm/refute hypotheses, in which case it is also known as confirmatory factor analysis (Kim & Muller, 1978b).

Since there are different uses of factor analysis, the exact relationship obtained from the variables may not be clear. Moreover, the types of strategies utilized to conduct factor analyses can be utilized in different ways simultaneously. For example, half of the variables can be used in exploratory strategies, while the other half can be used in testing hypotheses developed from the exploration of the first half of the variables (Kim & Muller, 1978a). Thus, there are numerous opportunities for use within the technique of factor analysis.

Unsurprisingly, conceptually, factor analysis is easy to learn. Practical applications of factor analysis, by contrast, are not so simple due to the sheer number of strategies that can be used to analyze variables. As a result, the underlying problem related to factor analysis is that variables tend to be much more complex than the analysis shows. For instance, factor analysis assumes that measurement levels will match the requirements of the analysis. However, some model components may be unrealistic in relation to the data provided, and minor factors may have a much more significant impact on the overall results than expected (Kim & Muller, 1978b). Therefore, researchers must be prepared to make unbiased decisions in relation to assumptions for the analysis and must be aware that the varying strategies can yield different final results.

The principal components (PC) approach was used in this study. Briefly, PC proceeds to solve the characteristic equation for the variance-covariance matrix (the basis for the correlation matrix), $[\mathbf{S} - \lambda_i \mathbf{I}] \mathbf{b}_i = \mathbf{0}$. The solution ultimately yields *eigenvalues* or λ_s , and *eigenvectors* or b_s . A given λ represents the variance accounted for by a given factor, and there will be as many λ_s as there are factors, as there are

variables. The b s, with one b for each variable on each factor, are weights for combining the variables and producing scores on the factors. Sets of *factor loadings or structure coefficients*, or F s, with as many F s as there are factors and within a given F , as many f s as there are variables, can be obtained by dividing a given b by the square root of its corresponding λ , or, of course by correlating the original variables with a given factor score. For example, given three variables, A, B, and C, and three Factors I, II, and III, obtain the factor loadings for Factor I by multiplying each weight in b_I by the square root of λ_I . *Factor score coefficients* are obtained for a variable by multiplying that variable's factor loading by the inverse of that factor's eigenvalue.

Multidimensional Scaling

Multidimensional Scaling (MDS) is used because of its suitability in standard data analysis. It attempts to organize objects in space with a specific number of dimensions in order to reproduce observed dissimilarities or distances in data (Borg, Groenen, & Mair, 2013; Groenen & Borg, 2014). This makes it possible to explain dissimilarities or distances between results with regard to the underlying dimensions. Similar to factor analysis, in MDS, axes are oriented, and so in the final solution, the actual axis orientation is arbitrary. For example, in using matrices of the distances between two points on a road from a road-map, the matrix can be analyzed with the aim of reproducing the distances on the map in two dimensions.

Using the MDS analysis will likely involve using a two dimensional representation for the two points locations, creating a two-dimensional map Borg, Groenen, & Mair, 2013; Groenen & Borg, 2014. In whatever direction the map is rotated, the distances between the two points do not change, although the view of

them may change. In the final analysis, a researcher may orient the axes of the distances due to her or his subjective viewpoints and decisions as the researcher selects the orientation that is easiest to explain such as an East West or North South orientation. Porter (2014) used the approach of MDS to see the differences that may exist among players in baseball. This research provides an overview of how to use and apply the mathematical concept of MDS in a study.

MDS is a method for rearranging points in space in a manner that is efficient so that a point that best estimates the observed distances can be arrived at by moving around objects in the defined space as dictated by the number of dimensions and tests the accuracy with which the distances between points can be reproduced using a new configuration (Borg, Groenen, & Mair, 2013; Groenen & Borg, 2014). MDS, therefore, implements an algorithm for minimizing functions that analyzes different configurations with the aim of maximizing the quality of fit or minimizing a lack of fit. The MDS algorithm used in this thesis was Kruskal and Wish's ALSCAL (Kruskal & Wish, 1978). To measure the quality of fit, the concept of stress was used. The stress measure accurately estimates the goodness of a fit for the observed distance matrices. The Phi (the raw stress value) is defined by the relationship:

$$\text{Phi} = \sum [d_{ij} - f(\delta_{ij})]^2$$

Where d_{ij} refers to the distances that have been reproduced based on the numbers of dimensions, while δ_{ij} is the data input or the observed distances. $f(\delta_{ij})$ denotes a non-metric transformation monotone of the observed distances. It is used for the purpose of reproducing the common rank-ordering of observations in the input data during analysis. The Shepard diagram is also a useful tool for measuring

quality of fit; this is achieved by plotting the reproduced distances for a specific number of dimensions against distances (observed input data). The scatter plot produces a Shepard diagram that is usually negatively sloped and shows the step function. The negatively sloped line represents \hat{D} values, which is the outcome of $f(\delta_{ij})$, the monotone transformation for input data (Borg, Groenen & Mair, 2013). Using more dimensions results in a better fit for the observed results because the stress is smaller.

Both PC and a two-dimensional MDS analysis were applied to statistics test items which assessed a skill or knowledge as identified below:

- Q1 Interpreting a histogram
- Q2 Identifying measurement levels
- Q3 Find/compute a median score
- Q4 Compute a standard deviation
- Q5 Find a percentile using the Standard Normal Distribution (SND)
- Q6 Compute and interpret Z-scores
- Q7 Find P-values for means using the SND
- Q8 Compute and interpret a confidence interval using the SND
- Q9 Compute the sample size required for a specifically-sized margin of error
- Q10 Compute and interpret a confidence interval using a t-distribution
- Q11 Perform a hypothesis test using the normal curve
- Q12 Give a practical definition of statistical power
- Q13 Identify and instance of a Type I Error
- Q14 Identify and instance of a Type II Error
- Q15 Interpret the results of hypothesis test using a t-distribution
- Q16 Identify the Law of Large Numbers
- Q17 Define r-squared
- Q18 Compute a predicted value of Y given a value of X
- Q19 Match a value of r_{XY} with a scatterplot
- Q20 Match a value of r_{XY} with a scatterplot

CHAPTER IV

RESULTS

The items were submitted to principal components analysis with both oblique and orthogonal rotations of the initial structure. The oblique components were not well correlated so the orthogonal result was retained for interpretation. Components associated with eigenvalues of 1.0 or larger were retained, and interpretations were based on items that demonstrated loadings of ± 0.40 or larger. The first two components appear stable and relevant, and reflect the major purposes of the course. The third through fifth factors are basically about definitions, but are not well marked and can be discarded, especially as the third and fifth structures contained only two items each (see Table 1).

Table 1.
Varimax Rotation of Components

	Component					h ²
	General Knowledge	Inference	Defining errors	Defining Terms	Defining Terms	
Q1 Interpreting a histogram						.239
Q2 Identifying measurement levels					.675	.510
Q3 Find/compute a median score	.512					.415
Q4 Compute a standard deviation	.556					.319
Q5 Find a percentile using the Standard Normal Distribution (SND)	.497					.327
Q6 Compute and interpret Z-scores	.402					.263
Q7 Find P-values for means using the SND						.238
Q8 Compute and interpret a confidence interval using the SND		.546				.388
Q9 Compute the sample size required for a specifically-sized margin of error						.284
Q10 Compute and interpret a confidence interval using a t-distribution		.629				.422
Q11 Perform a hypothesis test using the normal curve		.495				.340
Q12 Give a practical definition of statistical power				.714		.549
Q13 Identify and instance of a Type I Error			.839			.744
Q14 Identify and instance of a Type II Error			.855			.742
Q15 Interpret the results of hypothesis test using a t-distribution		.552				.323
Q16 Identify the Law of Large Numbers				.476		.407
Q17 Define r-squared	.409			.420		.345
Q18 Compute a predicted value of Y given a value of X	.592					.355
Q19 Match a value of r_{XY} with a scatterplot					.639	.522
Q20 Match a value of r_{XY} with a scatterplot	.553					.451
Eigenvalue	3.57	1.39	1.11	1.06	1.05	
Percent Variance	17.87	6.95	5.55	5.32	5.22	

The items were further submitted to multidimensional scaling by the ALSCAL method for binary data. Two dimensions were extracted and the stimulus pattern appeared to present four neighborhoods that could be named in meaningful ways. The analysis provided more details than did the components approach. Table 2 displays the computed dimension coordinates. A plot of the values per dimension across stimuli (items) suggests that the dimensions differ mainly for items 10, 11, 12, 13, 14, and 15 (see Figure 1). These items concern inference and appear in sequence in Form A of the test. The same items are in a randomized order in Form B of the test.

Table 2.

MDS Coordinate Values

	Dimension One	Dimension Two
Q1 Interpreting a histogram	0.357	0.143
Q2 Identifying measurement levels	0.188	0.075
Q3 Find/compute a median score	0.175	0.091
Q4 Compute a standard deviation	0.330	0.172
Q5 Find a percentile using the Standard Normal Distribution (SND)	0.399	0.207
Q6 Compute and interpret Z-scores	0.364	0.189
Q7 Find P-values for means using the SND	0.282	0.147
Q8 Compute and interpret a confidence interval using the SND	0.388	0.202
Q9 Compute the sample size required for a specifically-sized margin of error	0.398	0.207
Q10 Compute and interpret a confidence interval using a t-distribution	0.229	0.284
Q11 Perform a hypothesis test using the normal curve	0.259	0.321
Q12 Give a practical definition of statistical power	0.236	0.293
Q13 Identify and instance of a Type I Error	0.340	0.422
Q14 Identify and instance of a Type II Error	0.269	0.334
Q15 Interpret the results of hypothesis test using a t-distribution	0.211	0.262
Q16 Identify the Law of Large Numbers	0.360	0.144
Q17 Define r-squared	0.315	0.126
Q18 Compute a predicted value of Y given a value of X	0.311	0.124
Q19 Match a value of r_{XY} with a scatterplot	0.205	0.082
Q20 Match a value of r_{XY} with a scatterplot	0.498	0.199

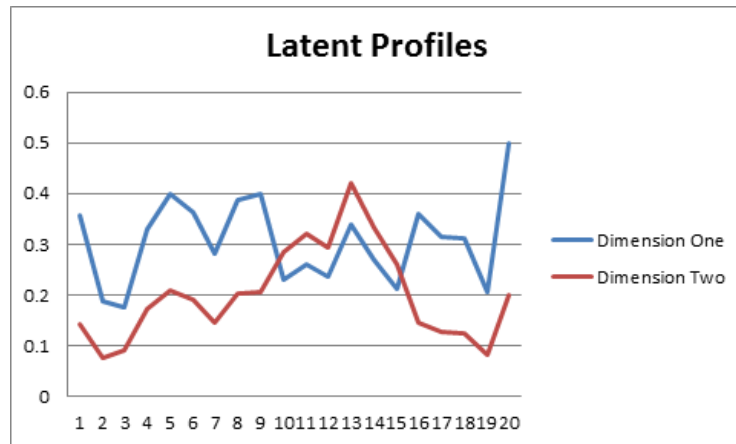


Figure 1. Profile plots for two dimensions of test items

The customary two-dimension plot (see Figure 2) suggested three meaningful interpretations: General knowledge and computation skills; Interpretation skills; and Giving definitions. A fourth neighborhood contains items which could not be interpreted in a meaningful way. The definitions neighborhood or quadrant incorporated many definition tasks/items and defined a stable construct as compared with the PC results that would lead to the discard of some definitions items.

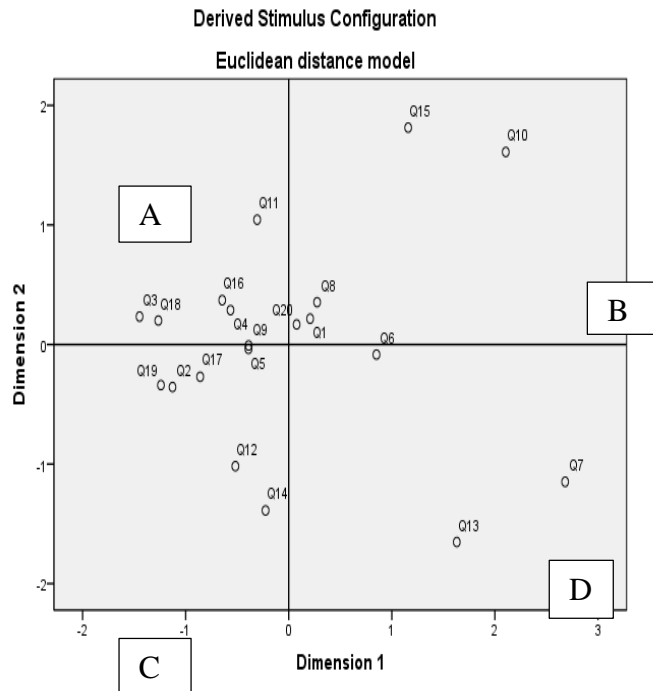


Figure 2. MDS of the exam items. Note: A: General knowledge with computing and interpreting, B: Interpretations, C: Definitions, D: Undefined, hybrid

Legend:

- Q1 Interpreting a histogram
- Q2 Identifying measurement levels
- Q3 Find/compute a median score
- Q4 Compute a standard deviation
- Q5 Find a percentile using the Standard Normal Distribution (SND)
- Q6 Compute and interpret Z-scores
- Q7 Find P-values for means using the SND
- Q8 Compute and interpret a confidence interval using the SND
- Q9 Compute the sample size required for a specifically-sized margin of error
- Q10 Compute and interpret a confidence interval using a t-distribution
- Q11 Perform a hypothesis test using the normal curve
- Q12 Give a practical definition of statistical power
- Q13 Identify and instance of a Type I Error
- Q14 Identify and instance of a Type II Error
- Q15 Interpret the results of hypothesis test using a t-distribution
- Q16 Identify the Law of Large Numbers
- Q17 Define r-squared
- Q18 Compute a predicted value of Y given a value of X
- Q19 Match a value of r_{XY} with a scatterplot
- Q20 Match a value of r_{XY} with a scatterplot

Reliability Analysis

The items were submitted to reliability analysis by Cronbach's internal consistency approach. Equations for Cronbach's Coefficient Alpha reliability coefficient are as follows. Based on sample data: $\frac{k}{k-1} \left(1 - \frac{\sum S_i^2}{S_T^2} \right)$.

In words, get the variance of each item and sum the item variances. Divide the sum of item variances by the total test variance. Subtract that quotient from 1.0 and multiply the difference by the number of items (k) over the number of items less 1.0.

The standardized version uses the average inter-item correlation: $\frac{k \overline{\text{Corr}}}{1+(k-1) \overline{\text{Corr}}}$

where $\overline{\text{Corr}}$ is $\frac{2}{k(k-1)} \sum_{i < j} \text{Corr}(i, j)$.

In words, manipulate the average inter-item correlation as a function of number of items (k). The average inter-item correlation is of all non-redundant and non-trivial correlations among the items (trivial is an item correlated with itself, while redundant would be item Y with item X, when item X with item Y has already been processed.) To do this, find the sum of all relevant correlations and then manipulate the sum as a function of k.

Using standard criteria for gauging reliability coefficients (Aiken, 1985), overall reliability is adequate, and no items appear to harm the coefficient by their presence (see Table 3.) The item analysis supported the usefulness of the test and did not provide evidence that the test items should be amended.

Table 3.

Alpha Reliability for the Examination

Reliability Statistics		
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.742	.744	20

	Mean	Minimum	Maximum	Range
Item Means	.655	.431	.908	.477
Inter-Item Correlations	.127	.017	.528	.511

Final exam test questions	Correlate Item VS. Total	α if Item Deleted
Q1 Interpreting a histogram	.357	.727
Q2 Identifying measurement levels	.188	.741
Q3 Find/compute a median score	.175	.740
Q4 Compute a standard deviation	.330	.730
Q5 Find a percentile using the Standard Normal Distribution (SND)	.399	.725
Q6 Compute and interpret Z-scores	.364	.727
Q7 Find P-values for means using the SND	.282	.734
Q8 Compute and interpret a confidence interval using the SND	.388	.725
Q9 Compute the sample size required for a specifically-sized margin of error	.398	.725
Q10 Compute and interpret a confidence interval using a t-distribution	.229	.738
Q11 Perform a hypothesis test using the normal curve	.259	.736
Q12 Give a practical definition of statistical power	.236	.737
Q13 Identify and instance of a Type I Error	.340	.729
Q14 Identify and instance of a Type II Error	.269	.735
Q15 Interpret the results of hypothesis test using a t-distribution	.211	.740
Q16 Identify the Law of Large Numbers	.360	.728
Q17 Define r-squared	.315	.731
Q18 Compute a predicted value of Y given a value of X	.311	.733
Q19 Match a value of r_{XY} with a scatterplot	.205	.739
Q20 Match a value of r_{XY} with a scatterplot	.498	.716

Standard Item Analysis

Items were analyzed by the conventional item analysis procedures.

Discrimination indices are given below, using the formula

$$\text{Index (DI)} = \frac{n_{\text{upper}} - n_{\text{lower}}}{n_{\text{neither}}}$$

where “n upper” and “n lower” are the frequencies in the upper and lower quartiles of students answering an item correctly, and “n either” is the sample size in either quartile. The Index can also be computed as the difference in the percentages of students passing an item between the upper and lower quartiles or $\text{DI} = \text{Upper\%} - \text{Lower\%}$.

DI ranges from 0.00 to 1.00, with adequate values in the mid-range and higher (Lin & Gronlund, 1995). Most of the DI values were about 0.50; the few that were smaller marked items that would make for “basement” or “easy” items that measure more rudimentary skills and knowledge that most students will have attained (see Table 4).

Table 4.

Discrimination of Items

Item	%Upper	%Lower	DI
1	92.1	26.3	65.8
2	89.1	51.3	37.8
3	98.3	77.1	21.2
4	96.1	41.2	54.9
5	97.2	38.7	58.5
6	88.5	21.1	67.4
7	75.9	19.7	56.2
8	93.6	25.9	67.7
9	99.1	35.2	63.9
10	80.8	29.7	51.1
11	89.8	34.8	55.0
12	88.0	35.9	52.1
13	84.0	17.5	66.5
14	90.0	32.3	57.7
15	83.5	37.8	45.7
16	96.2	42.5	53.7
17	85.5	30.3	55.2
18	99.4	67.3	32.1
19	91.7	50.8	40.9
20	97.4	22.7	74.7

Note: n = 532 in each quartile group.

Distractor Analysis and Item Difficulty

Inspection of response frequency across the four alternatives per item suggests that the distractors are about equally effective in evoking incorrect responses and that the items tend to be of appropriate difficulty (see Tables 5 and 6). A few items may be too difficult (with indices lower than 50 percent) due to lack of instruction on the topics or item structure itself, and several can be seen as “basement” items or those that nearly all students answer correctly. Note that the percentages of students who

answered an item correctly is generally in the acceptable range of 40 to 80 percent for a measure of item difficulty (Lin et al., 1995), across the set of 20 items.

Two forms of the test exist, differing only in order of item presentation (the items are identical between forms). Comparison of students taking either Form A or Form B suggests about equal achievement regardless of form.

Table 5.

Item Performance: Distractors and Difficulty, Form A

Item	A	B	C	D
1	6.5	10.5	22.8	<u>60.2</u>
2	21.6	2	<u>72.4</u>	4.1
3	<u>89.1</u>	5.3	5.5	0.1
4	8.4	15.4	<u>74.3</u>	1.9
5	3.8	15.0	<u>75.7</u>	5.5
6	3.2	4.5	35.7	<u>56.6</u>
7	<u>45.6</u>	39.8	9.8	4.8
8	11.5	<u>64.7</u>	16.8	6.9
9	10.0	9.6	<u>74.5</u>	5.9
10	<u>51.6</u>	2.6	3.2	42.6
11	6.9	27.6	<u>62.3</u>	3.2
12	22.4	8.7	8.3	<u>60.7</u>
13	40.7	<u>48.6</u>	4.3	6.4
14	5.6	5.7	29.2	<u>59.5</u>
15	34.4	10.9	2.7	<u>52.0</u>
16	10.0	7.1	7.7	<u>75.3</u>
17	13.1	7.8	<u>65.7</u>	13.4
18	2.1	1.5	<u>90.3</u>	6.1
19	<u>76.7</u>	1.6	19.6	1.9
20	<u>73.0</u>	13.8	7.2	4.4

Note: Entries are percentages of student choosing an alternative on a given item. Percentages for correct answers are underlined and italicized in boldface.

Table 6.

Item Performance: Distractors and Difficulty, Form B

Item	A	B	C	D
1	6.5	11.8	24.9	<u>56.8</u>
2	27.4	1.4	<u>62.4</u>	4.0
3	<u>91.6</u>	4.9	3.3	0.2
4	10.3	11.5	<u>75.6</u>	2.7
5	5.0	13.6	<u>75.9</u>	5.5
6	3.9	3.8	40.1	<u>52.2</u>
7	<u>38.9</u>	48.2	8.0	4.9
8	17.0	<u>59.9</u>	15.4	7.7
9	9.6	12.4	<u>72.2</u>	5.8
10	<u>43.3</u>	2.0	4.7	50.0
11	6.9	33.8	<u>56.0</u>	3.3
12	22.3	8.6	5.9	<u>62.2</u>
13	43.9	<u>45.6</u>	5.1	5.4
14	6.9	6.4	26.7	<u>60.0</u>
15	32.1	10.7	2.1	<u>55.0</u>
16	9.9	6.0	5.8	<u>78.2</u>
17	14.7	9.4	<u>63.6</u>	12.3
18	2.2	1.1	<u>91.7</u>	5.0
19	<u>67.9</u>	2.3	26.6	1.8
20	<u>65.9</u>	17.4	11.4	5.3

Note: Entries are percentages of student choosing an alternative on a given item. Percentages for correct answers are underlined and italicized in boldface.

CHAPTER V

CONCLUSION

Making a legitimate decision regarding a strategy for a particular examination requires the learning of the suppositions, restrictions, and data used in measuring an achievement profile. In addition to standard item performance investigations as done here, the current study attempted to identify profiles of test items that combined in meaningful ways while preserving the identities of the items themselves, within a single test instead of test performance within a battery of examinations.

Among the methods used here to study the structure of the final exam, the MDS model has the advantages of being effortlessly connected to tests of any size, grouping individuals on a continuum scale, and utilizing individual profile lists for further theory studies; however, a few provisos should be noted. To begin with, the determination of the number of measured versus latent or hypothetical variables impacts interpretability and reproducibility of results. Second, the relative importance of the measurements is to some degree not easily determined. A few scientists have utilized a bootstrapping technique to gauge the presumed and demonstrable importance of scale qualities (Jaworska & Chupetlovska-Anastasova, 2009). Third, it is not well understood to what degree the profiles or dimensions found in MDS can be compared over populations; for example, when comparing women and men (Petroski, 2005). More investigation is needed in these areas. Statistical tests and statistical conclusions need to be reliable; reliability or unwavering quality is the general

constancy of a measure. By comparison, researchers seek to be confident in the conclusions drawn and results offered as possible states of nature.

This paper therefore proposes that research involving quantitative measures should be evaluated using various statistical measures that give various perspectives on the data and the research questions. We might be able to draw valuable generalizations based on many statistical points of view.

To make research inferences valid and accurate, researchers should ensure the tools used have reliability in terms of test-retest aspects to ensure that test scores are consistent if different test administrations are used. Often, when scientific studies and research are undertaken, the assumption is normally that if the same tests were repeated, the results would more or less be the same. However, this may not necessarily be true. We need to ensure reliability of test scores by evaluating the statistical models to be used using a mathematical analysis framework like MDS or factor analysis, which portray the ways in which various items combine or separate in demonstrating student achievement, and in defining the concept of achievement (Borg, Groenen, & Mair, 2013; Groenen & Borg, 2014).

Finally, the data gathered and analyses performed suggest that the test currently used has sufficient construct validity and reliability to warrant its further use. Correlations with course grades would provide sound evidence of concurrent validity and define the practical relevance of the test as a final measure of achievement in elementary statistics.

REFERENCES

- Aiken, L. R. (1985). *Psychological testing and measurement* (5th ed.) . Boston: Allyn and Bacon.
- Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika*, 80(2), 317-340.
- Borg, I., Groenen, P. J. F., & Mair, P. (2013). *Applied multidimensional scaling*. Heidelberg, NY: Springer.
- Groenen, P. J., & Borg, I. (2014). Past, Present, and Future of Multidimensional Scaling. In J. Blasius, & M. Greenacre (Eds.), *Visualization and verbalization of data* (pp. 95-117). Boca Raton, FL: CRC Press.
- Gronlund, N. E. (1968). *Constructing achievement tests*. Englewood Cliffs, New Jersey: Prentice-Hall.
- Jacobs, J., Hiebert, K., Givvin, H., Hollingsworth, H., Garnier, H., & Wearne, D. (2006). Does eighth-grade mathematics teaching in the United States align with the NCTM standards? Results from the TIMSS video studies. *Journal for Research in Mathematics Education*, 37, 5-32.
- Jaworska, N., & Chupetlovska-Anastasova, A. (2009). A review of multidimensional scaling (MDS) and its utility in various psychological domains. *Tutorials in Quantitative Methods for Psychology*, 5(1), 1-10.
- Johannsdottir, B. (2013). *The mathematical content knowledge of prospective teachers in Iceland* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 1368260169).

- Hill, H., Sleep, L., Lewis, J., & Ball, D. L. (2007). Assessing teachers' mathematical knowledge: What knowledge matters and what evidence counts? In F. J. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 111-155). Charlotte, NC: Information Age Publishing.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.
- Kim, J., & Mueller, C. W. (1978a). *Factor analysis: Statistical methods and practical issues*. Beverly Hills: CA: Sage.
- Kim, J., & Mueller, C. W. (1978b). *Introduction to factor analysis: What it is and how to do it*. Beverly Hills: CA: Sage.
- Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills: CA: SAGE.
- Lin, R. L. & Gronlund, N. E. (1995). *Measurement and evaluation in teaching* (7th ed.). Englewood Cliffs, N.J. : Merrill
- Macnab, D., & Payne, F. (2003). Beliefs, attitudes, and practices in mathematics teaching: Perception of Scottish primary school student teachers. *Journal of Education for Teaching*, 29(1), 55-68.
- Magnusson, D. (1966). *Test Theory*. Reading, Massachusetts: Addison-Wesley.
- Mewborn, D. (2003). Teaching, teachers' knowledge, and their professional development. In N. C. Mathematics, J. Kilpatrick, W. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 45-53). Reston, VA: The National Council of Teachers of Mathematics.

- NCTM, National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.
- Petroski, G. F. (2005). *Statistical tests in the DFIT framework: A Monte Carlo evaluation of conventional methods and a bootstrap alternative* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 3189946)
- Porter, K. (2014). *Analyzing trends in baseball hall of fame voting, through the use of multidimensional scaling* (Master thesis). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 1558059)
- Schmidt, W. H., Wang, H. C., & McKnight, C. C. (2005). Curriculum coherence: An examination of US mathematics and science content standards from an international perspective. *Journal of Curriculum Studies*, 37(5), 525-559.